Compositional hierarchical structure evolves through cultural transmission

Compositional hierarchical structure evolves through cultural transmission: an experimental study

Carmen Saldana, Simon Kirby, Rob Truswell and Kenny Smith
Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences,
University of Edinburgh

Abstract

Compositional hierarchical structure is a prerequisite for productive languages; it allows language learners to express and understand an infinity of meanings from finite sources (i.e., a lexicon and a grammar). Understanding how such structure evolved is central to evolutionary linguistics. Previous work combining artificial language learning and iterated learning techniques has shown how basic compositional structure can evolve from the trade-off between learnability and expressivity pressures at play in language transmission. In the present study we show, across two experiments, how the same mechanisms involved in the evolution of basic compositionality can also lead to the evolution of compositional hierarchical structure. We thus provide experimental evidence showing that cultural transmission allows advantages of compositional hierarchical structure in language learning and use to permeate language as a system of behaviour.

Keywords: iterated learning; artificial language learning; communication; compositionality; hierarchical structure

---

Correspondence concerning this article should be addressed to Carmen Saldana, Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, 3 Charles Street, Edinburgh, EH8 9AD. E-mail <carmen.saldana@ed.ac.uk>

1 Introduction

Productive compositional structure is a unique and universal characteristic of human language. This supports a remarkable capacity for generating an unbounded number of different linguistic signals to communicate complex meanings, predictable from the meanings of their constituent parts and the grammar according to which they are combined (Chomsky 1965; Hockett 1960; for reviews of compositionality, see Pagin and Westerståhl 2010; Szabó 2012).

How did this compositional structure evolve? Evolutionary linguists have effectively studied it as a product of cultural evolution (Brighton, Smith, & Kirby 2005; Christiansen & Chater 2008). Languages are culturally transmitted through a repeated cycle of learning and communicative interaction. These two aspects of cultural transmission impose interacting pressures that shape the evolution of linguistic structure: a pressure for learnability (for ease of acquisition) and a pressure for expressivity (for communicative effectiveness). Compositional structure allows language to be both expressive and learnable, allowing users to communicate potentially about anything "making infinite use of finite means" (i.e., a lexicon and a grammar) (Chomsky 1965).

Several experimental models of iterated learning have shown that basic compositional structure emerges from the trade-off between learnability and expressivity pressures in cultural evolution (Beckner, Pierrehumbert, & Hay 2017; Kirby, Cornish, & Smith 2008; Kirby, Tamariz, Cornish, & Smith 2015; Theisen-White, Kirby, & Oberlander 2011; Winters, Kirby, & Smith 2015); in these experiments, languages evolve in which simple forms (morphs), constructed by recombining reusable meaningless units (phones), map to simple meanings, and these forms further combine to create more complex meanings. However, these combinatorial processes can be reduced to simple concatenation. The resulting form-meaning mappings are therefore less rich than those in natural language, in that they do not contain any hierarchical constituent structure or sensitivity to word order.

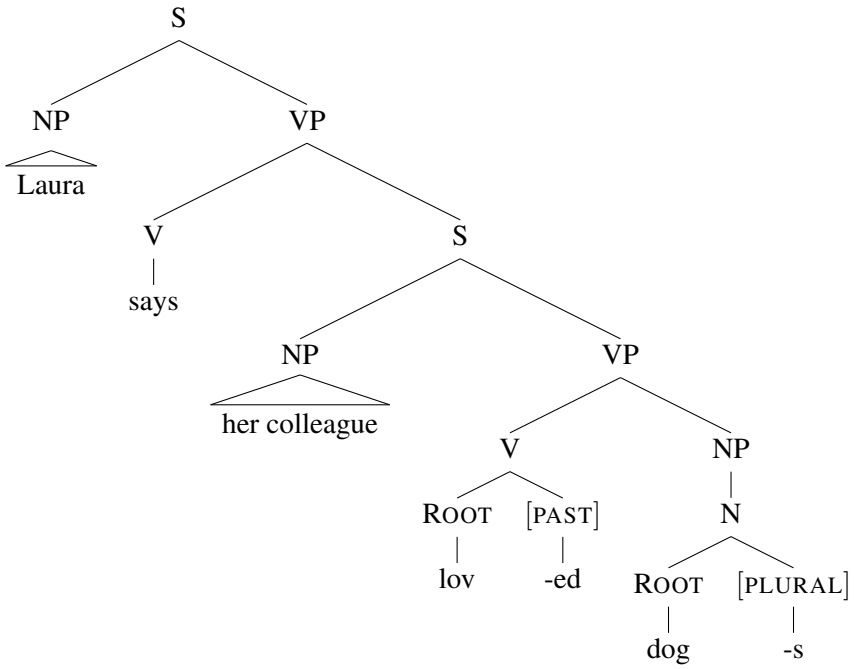
In this paper we show (by increasing the complexity of the meanings to be conveyed) that the same mechanisms at play in previous iterated learning experiments can lead to richer morphosyntactic structure, including both constituent structure and sensitivity to word order.

1.1 Compositional structure in natural languages

Sentences in natural language are organised into a hierarchy of constituents—known as constituent structure (Chomsky 1957), where each constituent higher in the hierarchy is built by recursively combining constituent units from lower levels. Sentences are built from other sentences and/or phrases; phrases, from other phrases and/or words; and words from other words and/or morphs. This hierarchical constituent structure provides systematicity in grammars: constituents are grouped into different syntactic categories whose members compose in definite and predictable ways with other linguistic material (Chomsky 1957; Pullum & Scholz 2007).

The structure of the sentence *Laura says her colleague loved dogs* in (1) illustrates these different levels of constituency.

(1)



In natural languages, compositionality is grounded in hierarchical constituent structure: meaning is determined relative to a structure, rather than to simple concatentation of forms. Form and meaning are isomorphic, that is, they have a structure-preserving one-to-one correspondence (Montague 1970). Isomorphism thus requires structure in form (i.e., morphosyntax) as well as in meaning (i.e., semantics). For example, in (1), the meaning of each non-terminal constituent node is composed of the meaning of its daughter nodes, which might themselves be complex: for instance, the meaning of *loved* is composed of the meaning of *love* and the past tense affix *-ed*, the meaning of *dogs* is derived from *dog* and the plural affix *-s*, and the meaning of *loved dogs* is composed of the meaning of *dogs* and *loved*. Further, the fact that (1) does not mean the same as *Laura says dogs loved her colleague* is also due to compositionality: if the same component parts are combined in a different structure, a different meaning results.

1.2 The cultural evolution of compositional structure

The intuition that language is compositional because compositionality facilitates language learning and use in communication is widespread, and studies in the field of mathematical linguistics ratify this intuition (Pagin 2012, 2013; Yang 2016). But how did compositional structure evolve? In order to establish a causal link between the observed structure in natural languages and its functional advantages, we need to explain how the advantages of compositional structure can permeate language as a system of behaviour shared at the population level (Brighton et al. 2005; Kirby 1999).

Various authors have argued that languages adapt over cultural time to maximise their learnability without jeopardising their expressivity (Brighton et al. 2005; Christiansen & Chater 2008; Regier, Kemp, & Kay 2015; Smith & Kirby 2012). Languages are learned from messy and relatively limited input but are nevertheless robustly transmitted in spite of this learning bottleneck: from limited data, language learners are able to acquire the necessary tools to generate novel interpretable expressions (Chomsky 1980; Kirby 2001). Because the poverty of the input presents

a challenge to the learner, language might adapt over time to maximise its learnability and reduce the impact of the bottleneck (Brighton et al. 2005; Kirby 2001; Zuidema 2003). There are several structural configurations that would solve this learnability problem (Kirby et al. 2008, 2015). A degenerate language in which all possible meanings are encoded by the same expression would be maximally learnable and could be transmitted intact; however, this learnability would be achieved through complete sacrifice of expressivity, since a degenerate language would not allow its users to discriminate between meanings. On the other extreme, a holistic language, in which each meaning to be conveyed is expressed by a distinct, idiomatic, unrelated expression would potentially be highly expressive. However, such a language could not survive transmission through a learning bottleneck: unless learners are exposed to the full language and their memory resources allow them to acquire it, a holistoc language cannot be transmitted intact from generation to generation. A compositional language resolves this tension between expressivity and learnability (Kirby et al. 2008, 2015). Provided that a learner's input data is sufficiently rich to allow the grammar and lexicon to be induced, compositional languages can be transmitted through a learning bottleneck but nonetheless allow for the production and interpretation of meaningful expressions.

In the last decade, evolutionary linguists have developed experimental models to study the emergence and evolution of linguistic structure in the laboratory (see Kirby, Griffiths, & Smith 2014; Scott-Phillips & Kirby 2010). Kirby et al. (2008) developed an Iterated Artificial Language Learning (IALL) paradigm to explore whether the pressures for learnability and expressivity previously described in computational and mathematical models of iterated learning (e.g. Brighton et al. 2005; Kirby & Hurford 2002) would have similar effects once idealised and rational learners were replaced with human participants. Kirby et al. (2008) showed the emergence of linguistic structure as languages were transmitted down generations of participants organised in transmission chains; with the implementation of a bottleneck in transmission (a pressure for learnability) and a filter to prevent ambiguous expressions from being transmitted to the next generation (a pressure for expressivity), holistic languages evolved to become compositional, a result replicated with a more adequate sample size by Beckner et al. (2017). Crucially, without this pressure for expressivity, languages only adapted to be learnable and evolved to be degenerate (see also Perfors & Navarro 2014; Silvey, Kirby, & Smith 2015). Further studies have shown similar effects with the introduction of communicative interaction (a natural promoter of expressivity) in transmission chains (Kirby et al. 2015; Theisen-White et al. 2011; Winters et al. 2015). These studies show that languages become structured as they are culturally transmitted through iterated learning and communicative interaction; critically, they also show that the same level of structure does not evolve from interaction alone.

1.3 Simple meaning spaces, basic compositionality

The meaning spaces utilised in previous IALL studies are very simple. For example, Kirby et al. (2008) used a meaning space that comprised 27 distinct meanings differentiated in three dimensions (see Figure 1), while the meaning space in Kirby et al. (2015) is even simpler. Consequently, the linguistic structure that emerged in these studies is correspondingly simple (see Figure 2). Although the structure is compositional in that the meaning of the expressions is derived from the meaning of the constituent parts, it lacks other aspects of compositional structure found in natural languages. For example, it lacks hierarchical constituent structure, which at a minimum would require the presence of complex expressions composed of complex expressions themselves. Moreover, as Galantucci and Garrod (2011) pointed out, IALL experiments have not yet shown the emer-

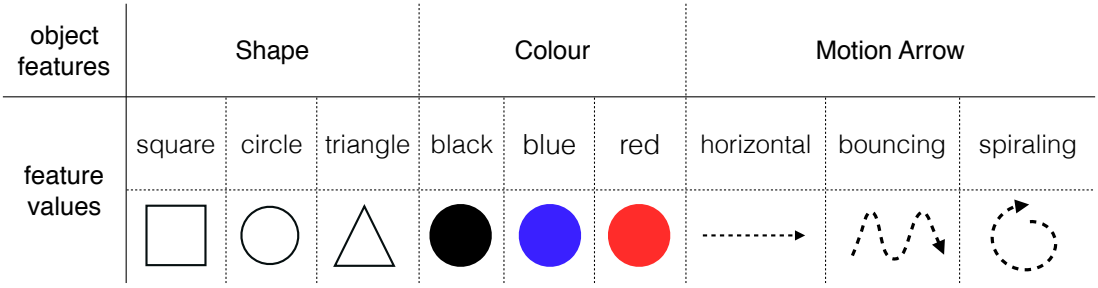


Figure 1. The visual stimuli used in Kirby et al. (2008) consisted of 27 pictures of coloured objects with arrows indicating motion. Each object feature (Shape, Colour and Motion Arrow) varied over three values: square, circle or triangle shape; black, blue or red colour; and arrows indicating horizontal, bouncing or spiralling motions.

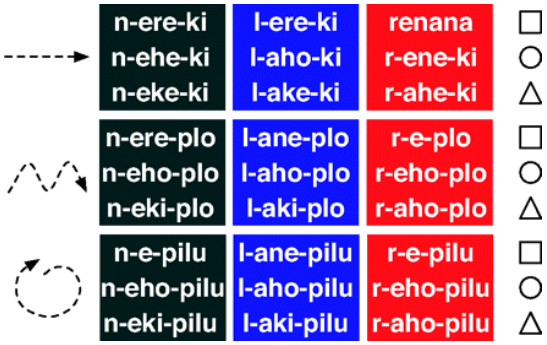


Figure 2. Example of a compositional language extracted from Kirby et al. (2008), Experiment 2.

gence of word order sensitivity, a type of compositionality the authors refer to as "positional", in which the same form takes on systematically different interpretations depending upon its position in the sequence, as discussed in section 1.1.

In the present study we aim to examine whether and how richer compositional structure evolves through cultural transmission in the laboratory. The guiding intuition is that the complexity of the compositional languages that evolve in IALL experiments is necessarily related to the complexity of the meaning space (i.e., the set of meanings participants learn and produce descriptions for), because a systematic mapping between semantic and morphosyntactic structure is a defining feature of productive compositionality.

We predict that, by introducing a more complex meaning space for speakers to learn and use, more complex linguistic structure will evolve in the same way basic compositionality has been shown to evolve in previous IALL studies with simpler meaning spaces. Our meaning space crucially includes motion events which involve two objects with different roles (focal and anchor), where each object can play each of the different roles. Each object has two properties (shape and number).

We predict that a meaning space with these characteristics will facilitate the emergence of

complex nominal elements encoding both shape and number, which can be depicted as nodes in constituent structure and thus comprise consistent syntactic categories. The emergence of complex nominal constituents provides two levels of complex constituency, the minimum required to show hierarchical constituent structure. Moreover, the need to distinguish objects according to their roles in the motion event requires the encoding of semantic roles in the argument structure either by means of morphology or word order rules. If word order rules emerge, we will be able to show the "positional" aspect of compositionality. On the other hand, if case marking systems emerge, it will also be the first time such functional morphology emerges in IALL studies (however, see van Trijp 2012). Altogether, this more complex meaning space offers the possibility for the evolution of complex nominal elements which are nodes in constituent structure and not mere object or property labels.

2 Experiment 1: transmission and artificial pressure against ambiguity

2.1 Method

The experiment utilises an Iterated Artificial Language Learning paradigm (Kirby et al. 2008, 2015). In overview, each participant in a transmission chain is trained on an artificial language based on some linguistic data, and then during testing produces a new set of linguistic data which will be the input data for the next learner in the chain. We ran four transmission chains of eight generations each. We implemented a strict artificial pressure against ambiguity during testing to block the evolution of degenerate languages.

2.1.1 Participants. Thirty-two participants were recruited to participate in an artificial language learning study through the University of Edinburgh's Careers Service database of student and graduate employment. All participants were native speakers of English (mean age 22 years, age range 18–42). Participants received a payment of £9. The experiment was conducted in accordance with the ethics procedures of Linguistics and English Language, The University of Edinburgh.

2.1.2 Stimuli. Participants were asked to learn, and then reproduce, an artificial language which provided descriptions for scenes of motion events. Motion events were represented using videos, descriptions were presented as text. We created 80 animated scenes to represent 80 motion events. Each video was five seconds long and featured one or more objects performing a motion; in some scenes this motion took place on a blank screen, in others the motion was relative to another object or set of objects.

More precisely, each scene featured a *focal* object or objects and, optionally, an *anchor* object or objects. There were two types of objects, squares and circles; each object could appear singly or as part of a group of multiple (i.e., nine) objects of the same shape (e.g., a group of nine circles or a group of nine squares). The focal object(s) in each scene performed one of two possible motions: sliding across the screen, or bouncing across the screen. That movement could occur once (resulting in a terminated motion event) or be continuously repeated for the entire duration of the scene (producing an ongoing motion event). If the scene featured anchor objects, the focal objects were initially on the opposite side of the screen from the anchor objects and moved towards the anchor objects; in scenes lacking anchor objects, the focal objects simply started on one side of the screen and moved to the other. The initial position of the focal objects (left or right side) was randomised on each presentation of each scene.

More formally, each motion event differed on five binary features: Shape of focal object, Number of focal objects, Motion, Aspect (terminated vs. ongoing), and Anchoring (whether the

event comprises an anchor object or not). Events with anchor objects(s) differed along two further binary features: Shape of the anchor object and Number of the anchor object—which contained the same features as Shape and Number of focal objects. This yields the full set of 80 possible motion events (16 events lacking anchor objects, 64 featuring anchor objects). Figure 3 provides a visualisation of the meaning features and values described.

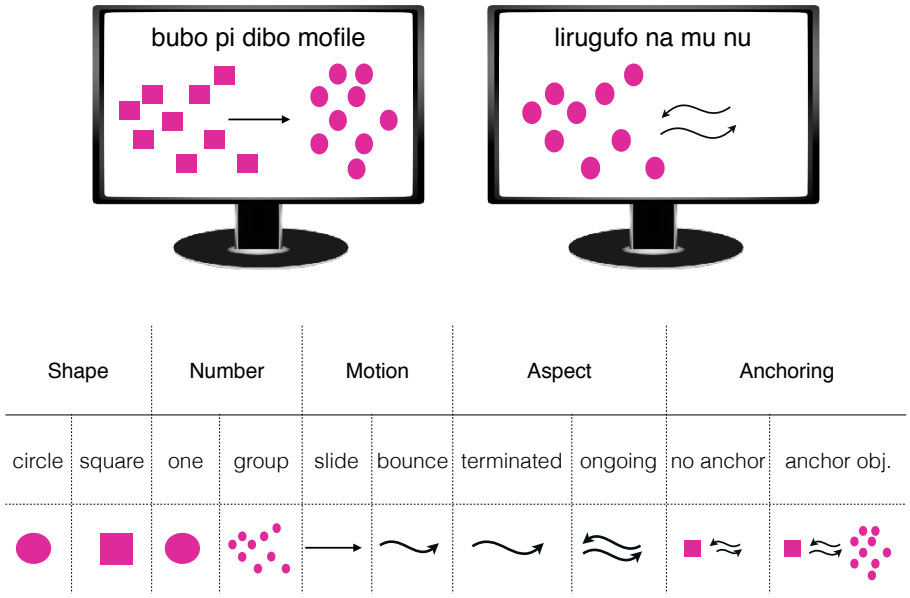


Figure 3. Features and values of the meaning space and example stimuli. The meaning space in this experiment consists of events which are composed of 5–7 different features (depending on the presence or absence of an anchor object respectively), each comprising two possible meaning values. The table (lower) shows the different features (collapsing Shape and Number in focal and anchor objects) and the different possible values for those features, with corresponding illustrations. Above, we show two examples of stimuli (with the motion represented with arrows) as they appear on screen during the learning phase: (left) a group of squares sliding towards a group of circles; (right) a group of circles bouncing back and forth, without any anchor object.

2.1.3 Initial languages. The initial languages to be learned by the first participant in each chain were a set of randomly generated holistic strings of lower-case letters, possibly including spaces. For each initial language, we generated 80 unique strings: each string consisted of 2–8 CV syllables, divided by spaces into 1–8 chunks (the number of chunks was randomly selected). These 80 strings were then paired randomly with the set of motion event scenes to create 80 scene-description pairs. We generated a separate initial language for each initial participant in each chain in order to eliminate any specific biases that might be imposed by the initial language.

2.1.4 Procedure.

Training and testing regime. Participants were asked to learn an artificial language made up of written labels for visual stimuli. They carried out the experiment at a computer terminal in isolated individual booths. All responses were entered using the keyboard. Participants received

written and verbal instructions before starting the experiment, and on screen at the start. The experiment was divided into two phases: a training phase and a testing phase.

During training, participants were taught a subset of 44 scene-description pairs from the total 80 pairs (randomly selected but always containing 3/4 of the non-anchored events and 1/2 of the anchored events). Each pair was presented three times in randomised order, yielding a total of 132 training trials, for a training phase duration of approximately 30 minutes. In each training trial, the description was shown in isolation for one second, then the associated scene was shown, accompanied by the description, for five seconds. After each presentation, participants underwent one of two recall tests (randomly chosen): retyping (50% of the time) or scene discrimination (other 50% of the time). In the retyping recall test, participants were presented with the motion event they had just seen, and were asked to retype its description. In the discrimination recall test participants were presented with two scenes side by side in randomised position, one of which they had just seen and the other selected randomly from the remaining 79 scenes in the meaning space; they were then asked to identify by button-press which of the two motion events matched the one they had just seen. These recall tests were intended to ensure that participants attended both to the training descriptions and their associated scenes.

During testing, participants were asked to describe all scenes twice in randomised order, yielding a total of 160 testing trials (approximate duration of 40 min). Note that participants were trained on 44 scenes of motion events but tested on all 80; this meant they were tested on events they had not been trained on. On each testing trial, the participant was presented with a scene for five seconds (this time without a description), and then asked to type its description in the artificial language. Participants were prompted to produce a different description whenever they entered a string which they had already used to describe a different scene during testing. This explicit demand for unique descriptions is intended to introduce a pressure against ambiguity to prevent the language from collapsing to a maximally-ambiguous single description, and is based on the method used by Verhoef (2012).

Transmission. Participants were organised into independent transmission chains, such that the language (set of scene-description pairs) produced by a participant at generation g is used as the training language for another participant at generation $g+1$ in that chain of transmission. Languages were formed by the set of descriptions participants last produced for each meaning. The initial participant in each chain, the first generation, is trained on a random target language, generated as described in section 2.1.3.

As mentioned in the description of the training and testing regime, we imposed a *bottleneck* on transmission. A language (either constructed with random strings or produced by a participant) consists of a set of 80 scene-description pairs. During transmission, we divide this into two subsets, a trained set (44 scenes, selected randomly) and an untrained set (the remaining 36 scenes) as described in section 2.1.4. This sub-setting procedure is implemented at each generation in a chain: the first participant is trained on a subset of the initial target language, subsequent participants are trained on a subset of the previous participant's output language. Participants were not informed of the source of the artificial language (i.e., that it was produced by another participant) until after completing the experiment.

2.2 Measures

2.2.1 Compositional structure: isomorphism between semantic and syntactic structure. Following Kirby et al. (2008, 2015), we quantify compositional structure as the z-score of

the Mantel Test between description similarities and scene similarities. Description similarity is calculated using normalised Levenshtein distance, i.e. the the number of characters that need to be changed, inserted or deleted to transform a description into another (Levenshtein 1966), divided by the length of the longest description (such that the maximum distance is 1). Scene similarity is calculated using Hamming distance (Hamming 1950), which is given by the number of feature values that are different between two scenes. Thus to quantify structure in a language, we first calculate the correlation coefficient between all pairs of edit distances in the set of descriptions and all pairs of edit distances of the corresponding scenes. This veridical (i.e. observed) coefficient gives us an indication of the extent to which similar meanings are associated with similar signals, as would be expected in a compositional language. We then calculate how likely the veridical coefficient between the two distance matrices is to appear by chance, using the Monte Carlo method of random sampling to produce a distribution of coefficients. For each language to be evaluated, we repeatedly randomise the associations between meanings and signals and re-calculate the correlation. We ran 10,000 samples, and from the distribution obtained, we extract the z-score for the veridical coefficient. If the z-score is greater than 1.645 (one-tailed)[1], we conclude that the veridical coefficient is unlikely to arise by chance ($p < 0.05$). High z-scores thus indicate a high degree of compositionality on this measure.

2.2.2 Reference. In order to minimise the influence of human biases in the linguistic analysis of the descriptions produced by participants, we extract form-meaning mappings automatically. We identify the referents of the lexical items[2] in the artificial languages by calculating the association strength between lexical items and the meaning feature values of scenes. We use Kendall's Tau-b rank correlation coefficient (Kendall 1938, 1945), which allows us to measure the strength and direction of the correlation between occurrences of a given lexical item and those of a given meaning feature value[3]. Values of Tau-b range from -1 to $+1$, indicating 100% negative or positive association respectively. A value of 0 indicates the absence of association, and the more distant Tau-b is from 0, the stronger the referential association. Thus the more a lexical item co-occurs with a specific feature value (e.g. a particular shape or movement), the higher Tau-b is.

2.2.3 Nominal syntactic categories. The emergence of complex nominal constituents (including at least two morphs encoding shape and number) is crucial to our study because they will provide the evidence required for hierarchical structure as well as for positional compositionality. We therefore need to evaluate whether morphs encoding shape appear adjacent to morphs encoding number and crucially, whether they constitute a syntactic category (i.e. have similar distributions).

[1] We use one-tail critical z-score values because we do not predict large negative z-scores; i.e., we do not expect significant negative correlations between description-similarities and scene-similarities. Moreover, as seen in the results section later on, z-scores obtained are very large and thus the use of one-tail instead of two-tail critical values does not alter the results.

[2] We consider lexical items those strings separated by spaces within a description.

[3] Before calculating Tau-b coefficients, we ran a diagnostic test to provide a more robust threshold for the significance of the dependence between lexical items and meaning feature values: we compute the mutual information of all pairs of lexical items and meaning feature values in a language. For each pair, we use the Monte Carlo method of random sampling to calculate how likely this veridical mutual information is to appear by chance. At each sample we randomise the mapping between scenes and descriptions and re-calculate the mutual information between the lexical item and the feature value. We run 10,000 samples, and from the resulting distribution, we calculate the probability to obtain by chance a mutual information equal or higher than the veridical; only if $p < 0.05$ we conclude a significant mutual dependency between a given pair (i.e., between a lexical item and meaning feature value) and proceed to calculate its Tau-b coefficient. Mutual information provides us with non-spurious correlations but not with a normalised value for the strength of the correlation or its direction (either positive or negative). Both strength and direction are thus obtained with Tau-b.

The syntactic category of a given grammatical unit can be inferred from the distributions in which it appears within sentences. The Taub-b measure described above allows us to identify morphs that refer to shape-objects. To determine whether these morphs form a syntactic category in the artificial languages, we evaluate their distributional similarity, on the assumption that members of a category will have similar distributions. We quantify the distributional properties of these constituents in a language (which we call nominals henceforth) as their set of backward transitional probabilities (BTP) (following McCauley & Christiansen 2011; Perruchet & Desaulty 2008) (i.e., the probabilities that each nominal has of being preceded by each of the lexical items in a language's lexicon). The BTPs for a given morph characterise its distributional properties; we then use the Jensen-Shannon distance metric (JSD) to measure the distances between all pairs of BTP distributions; the lower the distance between a pair of distributions, the more similar they are. The average JSD between all pairs then gives us an indication of the distributional similarity between all nominals. We then calculate how likely this veridical average JSD is to appear by chance using the Monte Carlo method of random sampling. At each sample, we randomly select a set of lexical tokens of the same cardinality of the set of nominals in the language and calculate the average JSD between all pairs of BTP distributions. We run 10,000 samples, and from the distribution obtained, we extract the z-score for the veridical average JSD. Low z-scores indicate short distances between BTP distributions of nominals and specifically, z-scores below -1.645 ($p < 0.05$, one tailed)[4] suggest that nominals within a language share similar distributions within sentences and thus constitute a syntactic category.

2.2.4 Order of nominal arguments. The same object or group of objects can appear in the focal or anchor roles. There are different ways in which a language describing the stimuli could mark nominal arguments for the semantic roles they perform: this could be morphologically encoded (e.g., via affixation, suppletive forms or functional particles), and/or cued by the order in which nominals appear within a sentence. For example, nominals appearing in first and second position in a sentence might systematically have focal and anchor semantic roles respectively; in this case, the position in which nominals appear could determine their meaning. We assess the systematicity of the order of nominal arguments by calculating the Shannon entropy of the different orders in a language.[5] A language without any defined pattern of the order of nominal arguments obtains the maximum possible entropy of 1 bit (i.e., a language with 50% anchor-focal and 50% focal-anchor orders, or a language with all undefined patterns), and a language with a consistent

[4]We use one-tail critical values because we do not predict large positive values. Large positive z-scores could only be obtained if the distribution between nominals were more dissimilar than obtained by chance. Given the conservativeness of the random sampling method (i.e, selection from tokens and not types, and descriptions are kept intact at each randomisation), we do not expect to obtain dissimilarity scores significantly distant from the mean of the random sample.

[5]Entropy measures how variable the order of nominal arguments is between sentences in a language. In an unambiguous compositional language there are only two possible nominal orders, either focal arguments precede anchor arguments (focal-anchor orders) or vice versa (anchor-focal orders). However, as linguistic structure emerges, the order of nominals may be undefined: participants will produce underspecified descriptions either because they might not have the lexicon to refer to objects and/or the number they appear in, or because they simply do not use the lexicon consistently. In order to measure the entropy of the system of nominal argument orders, we first exclude from the language those descriptions that refer to motion events which do not have an anchor object (16 descriptions) and those whose focal and anchor object are the same (a further 16 descriptions)—as the order of nominals there is not informative. We then calculate the frequency of focal-anchor, anchor-focal and undefined orders within the reduced language of 48 descriptions. Undefined orders introduce randomness into the system; in order to implement such randomness in the entropy measure, we split the frequency of undefined orders between focal-anchor and anchor-focal patterns equally. We then calculate the entropy of the resulting vector of frequencies.

order of nominal arguments would result in the minimum possible entropy, that is, 0 bits (i.e., a language with 100% anchor-focal or 100% focal-anchor orders).

2.3 Analysis and results

2.3.1 Qualitative analysis. For each language produced in Experiment 1, we extracted a matrix of associations between lexical items and their referents in the scenes as explained in section 2.2.2. With this matrix we were able to automatically gloss the meanings of the descriptions provided in the experiment and analyse their structure minimising any potentially biased interpretation of the semantics of a language. The induced dictionary for the final language A3 (where 3 indicates the chain number, and A indicates an *Artificial* pressure against ambiguity, differentiating these languages from those presented for Experiment 2) is shown as a matrix of associations in Figure 4 as an example. Examples of descriptions in the same language A3 with the corresponding glosses are provided in (2). In these descriptions, we observe that lexical items associated with shape-objects, which we call nominals, precede the lexical items associated with motion or aspect features in the scenes, which we refer to as verbal elements. Moreover, Language A3 uses word order as a morphosyntactic cue to interpret the different semantic roles of the nominal arguments in a sentence; the roles of focal and anchor object are consistently assigned to the first and second position in a description respectively. The correct interpretation of the semantic roles of each of the nominal arguments is crucial, as the same nominals can refer to focal and anchor objects. Language A3 also makes use of redundant functional markers of semantic roles and anchoring (i.e., the presence of an anchor object in the event) (see sentences 2a–b). One marker, *pifli*, systematically follows focal nominal arguments only in sentences that refer to scenes with more than one object; another marker, *trink/-i san/-s*, follows anchor nominal arguments. In addition, the form of the latter is conditioned by the number of the nominals: if one or more nominals are marked as plural, these anchoring markers will appear as *trinki sans* rather than *trink san* (see 2a-b).

(2) a. rons pifli mons trinki sans hula bu
square.group ANC [6] circle.group ANC.group slide terminated
'A group of squares slid towards a group of circles'

b. mon pifli ron trink san hula
square ANC circle ANC.one slide
'A circle slides towards a square back and forth'

c. mons hulai ai
circle.group bounce terminated
'A group of circles bounced'

Language A3 comprises two main categories: a nominal category that consists of complex constituents formed by morphs associated with Shape and Number features (the latter always suffixed to the former), and a verbal category formed by morphs associated with Motion, followed by a marker of Aspect in terminated events. Nominals are thus morphological complex lexical constituents with strict internal structure. Any linear word order rule has to respect the integrity of

[6] ANC stands for anchoring marker, particles that appear only with events that contain both a focal object and an anchor object.

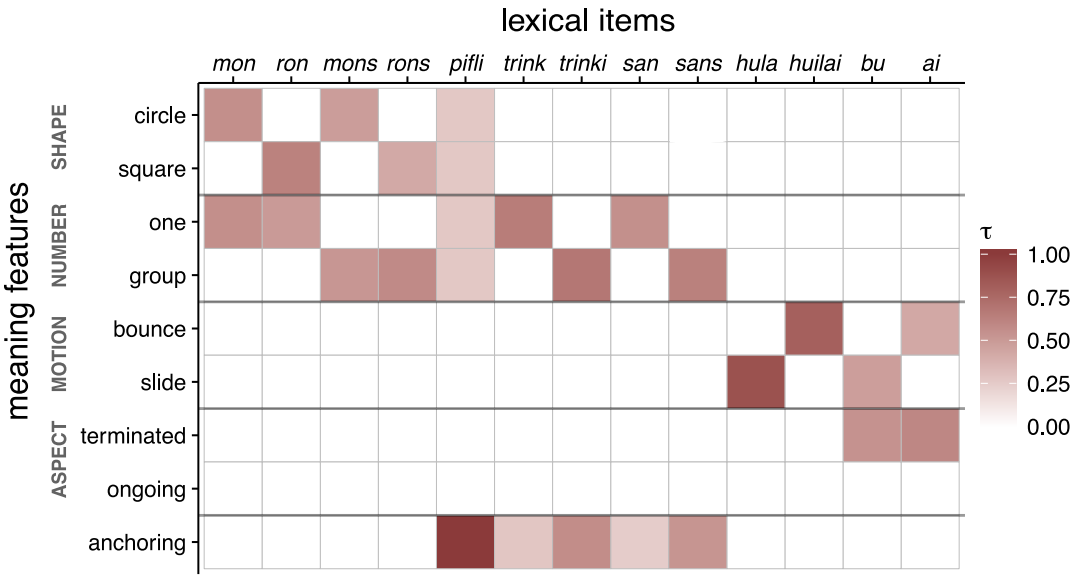


Figure 4. Heatmap illustrating the different semantic categories of lexical items (x axis) found in language A3 (chain 3) in relation to the meaning features they refer to (y axis). The heatmap scale represents the strength of the positive association between lexical items and meaning feature values (Tau-b coefficient, see section 2.2). We can distinguish three salient patterns in A3's lexicon which correspond to three different categories that we will call nominal, functional and verbal elements. Moving from left to right along the x axis in Figure 4 we find: lexical items associated with Shape (*circle* or *square*) and Number (*one* or *group*) which form a nominal category, a set of lexical items associated with Anchoring (presence or absence of an anchor object) as well as Number which form a functional category, and lexical items associated with Motion (*slide* or *bounce*) and/or Aspect (*terminated* or *ongoing*), which constitute a verbal category. In the nominal category we have *mon(s)* and *ron(s)*, which are the only items that refer to the shapes in the scenes. The affix *-s* acts as a plural marker and its absence marks singularity. Verbal elements are *huilai*, *hula* and *ai* and *bu*. Both *huilai* and *hula* are free morphs associated with Motion alone, and *ai* and *bu* act as their respective aspect morphemes (i.e., although separated by spaces, they cannot stand on their own), whose presence marks the events as terminated.

		A1	A2	A3	A4
circle	one	su**nyan**	p**ion**a/pijone	m**on**	cica/(cicaa)
	group	sa**nyan**	p**ion**dr-a/e	m**on**s	lumuse
square	one	vu**nyan**	f**ion**a	r**on**	demi/(dmei)
	group	va**nyan**	f**ion**dr-a/e	r**on**s	demi-toda/tofa/fora

Table 1
Nominals in the final languages: at generation 8 for languages A1–A3 and generation 7 for language A4. The elements in bold signal category markers, recurrent patterns across members of a lexical category. The parenthetical alternative items in A4 occur only as focal objects and only with sliding events.

		A1	A2	A3	A4
slide	ongoing	F tolo A (vale)	watashe, zu, yu, mebe	F A hula	F fumuse A
	terminated	F vero/velo A (re/te)	watashe, zu, yu, mebe	F A hula bu	F sahime A
bounce	ongoing	F galamete A (vale)	watashe, zu, yu, mebe	F A huilai	F fumuse A
	terminated	F vero/velo/galamete A te/re	watashe, zu, yu, mebe	F A huilai ai	F sahime A

Table 2
Verbal elements in the final languages: at generation 8 for languages A1–A3 and generation 7 for language A4. Elements in brackets are optional. The most frequent position of anchor nominal arguments is represented by A, and that of focal nominal arguments, by F. In language A2 there is system of verbal elements: the four lexical items listed appear randomly in sentences.

the nominals and cannot break them up, that is, number is always realised as (al least) a suffix in nominals.

Tables 1 and 2 below display the different lexical items under the nominal and verbal categories evolved in all languages (respectively). Languages A1–A3 are extracted from the final generations and language A4, from the penultimate generation—the participant in the last generation of chain 4 failed to learn the lexical items of the input language causing a drastic decrease in the language's structure (see section 2.3.2 to follow). Sentences mostly comply with the structures described in the two tables and therefore can be reconstructed from them. The order of nominal arguments is mainly fixed in A1, A3 and A4: focal nominal arguments precede anchors.

All languages encode Shape and Number within nominals. Moreover, we observe further sublexical structure in the morphs encoding Shape within nominals: most languages (A1–A3) contain a nominal category marker, signaled in bold in Table 1. Number is almost exclusively marked via affixation (simulfixation in A1 and suffixation in the rest) with the exception of the suppletive forms found in language A4 to mark the plurality of circle objects (see Table 1).

Whilst the encoding of Shape and Number is fairly systematic in the final generations across chains, the encoding of Motion and Aspect is not entirely (or at all) established in languages A1 and A2. Moreover, it is only in the last two generations that Motion starts to be encoded in language A4

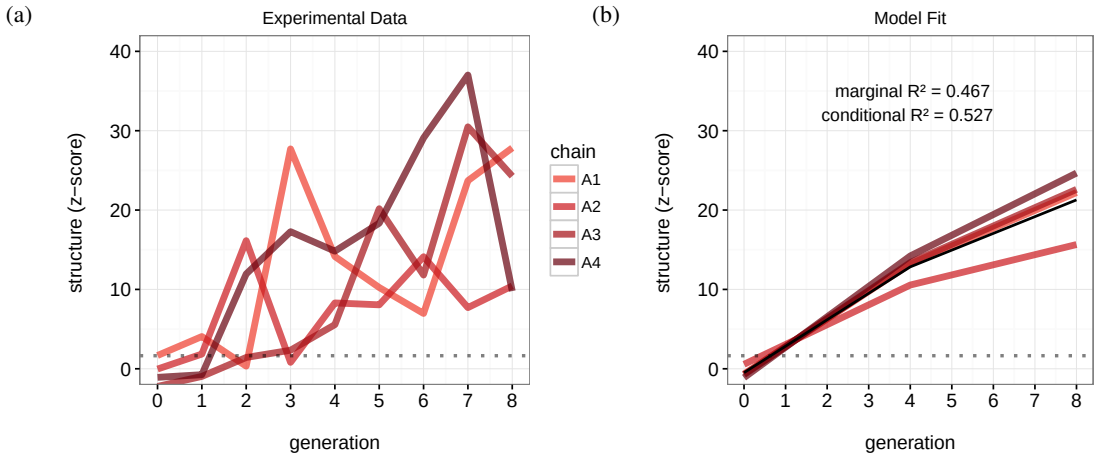


Figure 5. (a) Linguistic structure over generations for each of the four transmission chains. Linguistic structure increases as languages are transmitted through generations of learners. (b) Fitted values from the mixed-effects regression Model 1 for the four transmission chains and their average (in black). Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas their average in black represents the fixed effects estimates. In both plots, the dotted horizontal line represents the chance level (z-score 95%CI $= \pm 1.645$, one-tailed).

but it is not fully systematic either[7].

2.3.2 Compositional structure. We hypothesised that, by introducing a more complex meaning space for speakers to learn and communicate about, more complex linguistic structure would culturally evolve in the same way structure has been shown to evolve in previous IALL studies with simpler meaning spaces (Kirby et al. 2008, 2015). As discussed in section 2.3.1, linguistic structure indeed emerges to convey this complex meaning space through cultural evolution. Languages shift from holistic to compositional systems. Figure 5a shows the structure scores obtained in the experimental data (see measures in section 2.2.1). We observe that structure gradually increases as languages are transmitted through generations of participants; all languages are significantly structured from generation 4 onwards (chance level is represented by a dotted line in Figure 5a).

We used R (R Core Team 2000) and the package *lme4* (Bates, Mächler, Bolker, & Walker 2015) to perform a segmented linear mixed-effects model (SLMM) to explore the effect of generation on linguistic structure (measured as explained in section 2.2.1). We will call this Model 1. We ran a SLMM because it allows to easily quantify an abrupt change of the response function of a varying influential factor and we expect the influence of generation to be more distinct in the first generations and significantly less so in the latter generations as languages become structured. Unlike other types of growth curve modelling, SLMMs allow the identification of a specific point of change in an otherwise linear relation between generation and the dependent variable (structure), and most importantly, the direct effect of that point of change. In Model 1, Generation is partitioned into two intervals with one breakpoint at generation 4, and a separate line segment is fit to each inter-

[7]This late and sudden encoding of a previously underspecified meaning feature suggests that the participants might be employing a conscious strategy to increase expressivity.

	Generation							
	1	2	3	4	5	6	7	8
chain A1	NA	NA	1	1	1	1	1	1
chain A2	NA	1	NA	1	1	0.9	NA	1
chain A3	NA	NA	1	1	0.96	0.96	0.93	0.76
chain A4	NA	0.97	1	1	1	1	1	1

Table 3
Proportion of adjacent Shape and Number morphology for a specific object, either focal or anchor. Shape and Number are always encoded adjacent to each other in most languages from generation 2 onwards. The only case where we observe a notably lower proportion of adjacency is in the final generation of chain A3, where on top of Shape and Number morphology, case-like markers evolve which can agree in number with non-adjacent lexical items. Non-applicability (NA) in a given generation signals that adjacency cannot be measured due to the lack of Shape and/or Number morphology.

val[8]. In order to extract the best breakpoint we followed the procedure described in Baayen (2008, pp. 238–239): we fitted a series of models, one for each possible breakpoint in the range of generations, including breakpoints at generation 0 and 8, which equate to no breakpoint[9], then selected the breakpoint of the model with the lowest deviance, which was at generation 4.[10] As fixed effects, we entered Generation and an interaction between Generation and Indicator. As random effects, we introduced intercepts for Chain as well as by-Chain slopes for the effect of Generation. Figure 5b shows the predicted values based on the fixed and random parameter estimates obtained. We found a significant[11] effect of Generation ($\beta = 3.32, SE = 1.08, p = 0.005$), suggesting that structure increased as languages were transmitted down generations of learners. There was no significant interaction between Generation and Indicator ($\beta = -1.222, SE = 1.831, p = 0.509$), indicating that structure increased at approximately the same rate in the first and last four generations.

2.3.3 Compositional hierarchical structure: the emergence of complex (nominal) constituents within sentences. We hypothesised that complex constituents would evolve. In particular, we hypothesised that morphologically complex nominals which constitute a node in constituent structure could emerge, comprising at least morphs that refer to Shape and Number meaning features. Table 3 shows the relative frequency in which morphs referring to the Shape and Number features of a specific object (either focal or anchor) appear adjacent to each other within descrip-

[8] In order to introduce a breakpoint at generation 4, we first shifted the value of Generation so the intercept at 0 is in generation 4. We then introduce an Indicator variable that specifies whether or not each of the shifted values is greater than 0, that is, whether a generation belongs to the first or second segment.

[9] It is worth noting that this automatic procedure developed in Baayen (2008) yields a higher-than-nominal Type-I error rate of finding non-linearity. In order to check that multiple comparisons were not too problematic, we followed the simulation-based approach described in Vanhove (2014) to calibrate the p-values.

[10] Note that the breakpoint was extracted from the model comprising the data from experiments 1 and 2; we include it in this model for Experiment 1 alone even though it is not significant to assure consistency with Model 4, where the breakpoint is non-trivial.

[11] As in all models to follow, p-values were calculated using lmerTest (Kuznetsova, Brockhoff, & Christensen 2014). The library *lmerTest* calculates p-values of fixed effects from F statistics based on Satterthwate's approximation for denominator degrees of freedom, and it tests random effects using likelihood ratio.

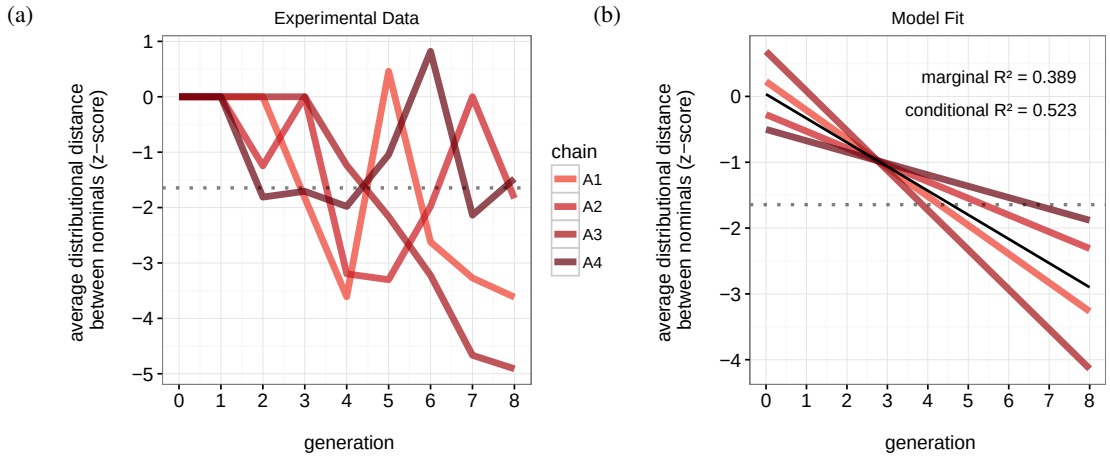


Figure 6. (a) Distributional distance between a language's nominals through generations for each of the four chains. The dotted line represents the lower bound of chance (z-score 95%CI $= \pm 1.645$, one-tailed); z-scores below it indicate that the distributional similarity between nominals is unlikely to arise by chance. Distributional distance between nominals decreases with generation as languages become more structured, suggesting the emergence of a nominal syntactic category. Nonetheless, only two languages stay consistently below chance from generation 6 onwards and only three end up below chance at the final generation 8. (b) Fitted values from the mixed-effects regression Model 2. Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas their average in black represents the fixed effects estimates.

tions. We find that they consistently appear adjacent to each other in most languages once morphs to encode Shape and Number evolve. Number morphs across languages are bound to Shape morphs via affixation. The only case where we find a notably lower proportion of adjacency is in the final generation of chain A3, where, as described in section 2.3.1, on top of the number morphology adjacent to shape morphology, long-distance number agreement evolves (i.e., encoded in redundant functional markers).

We now turn to test whether these complex nominal constituents can be syntactically categorised as nodes in the structure of yet more complex linguistic expressions (sentences). Syntactic categories are formed by constituents which arrange with other linguistic material in a similar way and thus share the same distributions in a sentence. We assessed the significance of the distributional distance between nominal constituents as explained in section 2.2.3. Figure 6a shows the z-scores of the average distance between the distribution of nominals in the descriptions of language. We observe the emergence of nominal syntactic categories across chains: the different nominals share similar distributions in the descriptions of a given language. Nevertheless, only two languages (chains A1 and A3) stay consistently below chance from generation 6 onwards, the other two languages (chains A2 and A4) are less stable and are distributed around the lower bound of chance ($z = -1.645$) by the final generation. Note that given the conservative nature of the analysis of distributional distance, which is carried out on un-annotated raw languages, if we obtain z-scores significantly below chance we should assume that the order of verbal elements in relation to nominals is mostly fixed. However, we cannot infer anything else about the order of nominals as this measure is blind to semantic roles (i.e., whether nominals refer to focal or anchor objects); we

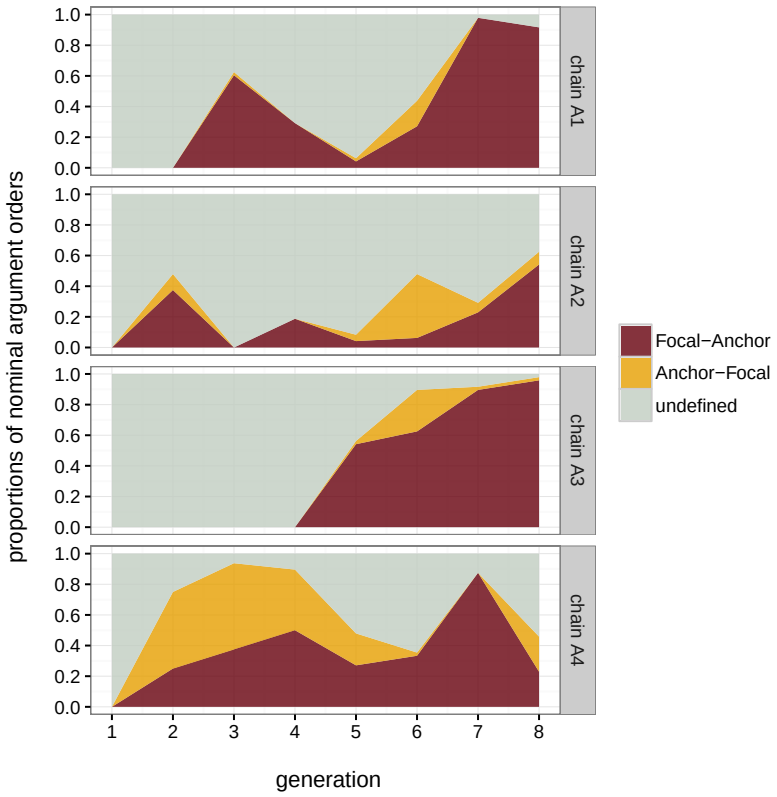


Figure 7. Proportion of word order types for nominal arguments by chain and generation. The proportion of undefined word orders decreases as languages become more systematic and in chains A1 and A3 we observe the evolution of a fixed Focal-Anchor word order towards the final languages. In chain A4, word order becomes systematically anchor-focal at generation 7 but de-systematises in the next generation as the participant fails to learn the vocabulary of the language. In the remaining chain A2, word order rules never evolve.

discuss the order of nominals within descriptions in relation to their semantic roles in section 2.3.4.

We performed a linear mixed-effects model, which we will call Model 2, to explore the relationship between the distributional distance of nominals and generation. We do not report a segmented model because the absence of a breakpoint constituted the best fit to the experimental data. We entered Generation as the only fixed effect (centred). As random effects, we introduced intercepts for Chain and by-Chain random slopes for the effect of Generation. Figure 6b shows the fitted values of Model 2 for fixed and random effects. Results showed a significant effect of Generation ($\beta = -0.367, SE = 0.120, p = 0.036$), suggesting that the distributions in which complex nominals appear do become more similar as languages are transmitted down generations of participants.

2.3.4 Word order rules for nominal arguments. We have previously mentioned that languages have various ways of encoding the semantic roles of the arguments in a sentence. Semantic roles can be encoded morphologically or can be cued by the position they occupy in a sentence. In section 2.3.1 we observed that all nominals could occupy both focal and anchor semantic roles, and only one language (i.e., A3) developed a morphological marker for the different semantic roles.

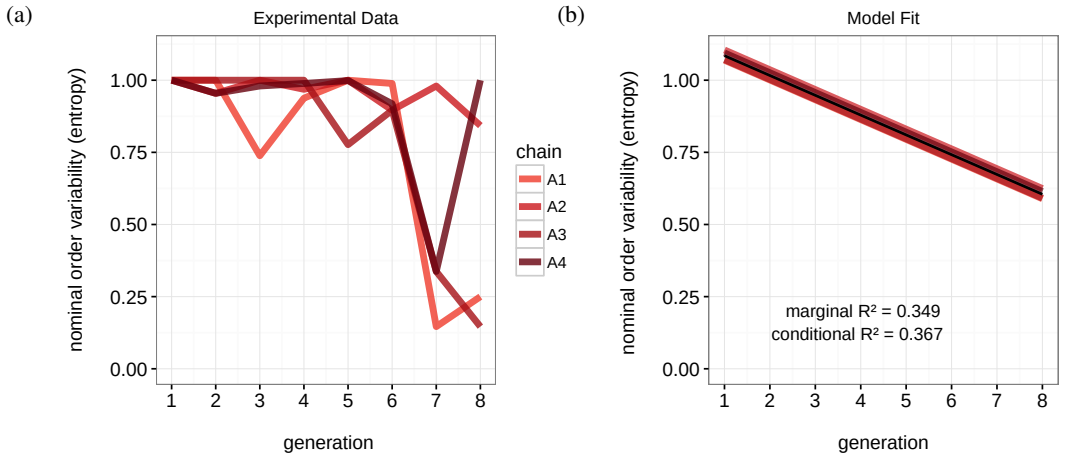


Figure 8. (a) Variability of nominal argument orders by generation and chain. (b) Fitted values from the mixed-effects regression Model 3 for the four transmission chains and their average (in black). Because Model 3 does not include random slopes, both coloured and black lines represent the fixed effects estimates.

The stacked area graphs in Figure 7 show the proportions of focal-anchor, anchor-focal and undefined orders at each generation for each of the four transmission chains in the experiment. A visual inspection of Figure 7 suggests that the proportion of undefined order decreases as languages are transmitted through generations. Moreover, we observe the evolution of a fixed order of nominal arguments in at least three out of the four chains (A1, A3 and A4).

We ran a linear mixed-effects model (Model 3) to test the effect of generation on the variability of nominal argument orders in a language—measured by the entropy of the system of orders as described in section 2.2.4. We entered Generation (centred) into the model as the only fixed effect; as random effects we enter intercepts for Chain[12]. Figure 8 shows the nominal order variability scores of the experimental data (Figure 8a) as well as the fitted values of Model 3 for fixed and random effects (Figure 8b). Results show a significant effect of Generation ($\beta = -0.06, SE = 0.017, p < 0.001$), suggesting that the order of nominal arguments becomes more consistent as languages are transmitted through generations of learners. Therefore, along with an increase of overall structure, the order of nominal arguments becomes more consistent suggesting the emergence of what Galantucci and Garrod (2011) called “positional" compositionality: the same exact nominal constituent can acquire different semantic roles (either focal or anchor) depending on its position.

3 Interim Discussion

In Experiment 1 we examined whether complex compositional structure would evolve in the same way basic compositionality has been shown to evolve in previous IALL studies (Kirby et al. 2008, 2015). We wanted to provide evidence for two properties of compositional structure found in natural languages that had not yet been shown in IALL studies: hierarchical constituent structure

[12]This was the maximum random effects structure allowed without convergence warnings. Moreover, the model with the inclusion of by-Chain random slopes for the effect of Generation was not significantly better ($\chi^2(2) = 0.597, p = 0.742$).

(i.e., complex constituents are built from further complex constituents), and argument structure whose semantic roles are marked via word order rules evidencing “positional" compositionality.

We showed that compositional structure evolved from holistic languages as they were transmitted through generations of participants. Languages developed morphology to match existing feature values in the meaning space establishing isomorphism between semantics and morphosyntax. These results replicate the results found in previous IALL studies (Kirby et al. 2008, 2015; Silvey et al. 2015) but with a more complex meaning space.

Our results further suggest the evolution of morphologically complex constituents which constitute a nominal syntactic category; they all share the same distribution within sentences and thus can be interchanged with each other to derive grammatical structures. Moreover, all nominal constituents within a given language share a morphological category marker and thus high string-similarity (for similar results, see Carr, Smith, Cornish, & Kirby 2016; Nowak & Baggio 2016). These nominal constituents combine with each other and verbal elements to form more complex linguistic expressions. Compared to previous IALL studies, this is the most productive structure hitherto shown to evolve. Here we show at least two levels of the hierarchy of constituent types in natural languages: morphs combine to form word-like forms and these further combine to form sentence-like structures. We can thus describe the nominal elements that evolved as nodes within hierarchical constituent structure and not just isolated referring expressions or labels which combine via concatenation. Additionally, we show the evolution of word order rules in argument structure to encode the semantic roles in motion events. As languages become more structured, the order of nominal constituents in a sentence also becomes more systematic: the position in which the same nominal constituents appear determines the argument they refer to. Nowak and Baggio (2016) similarly show the emergence of word order regularity in a multigenerational signalling game, but given that their objects can only appear either in subject or object position but not both, “positional” compositionality cannot be evidenced; our study further supports the emergence of word order regularity through cultural evolution and provides evidence for positional compositionality.

Nevertheless, linguistic structure only fully evolved in two of our four chains: whilst nominal constituents encoding the shape and the number of the objects evolved in all languages, verbal constituents that systematically mapped to the motion and aspect features of scenes (which occur less frequently) only evolved in two. It is possible that the restrictiveness of the pressure for expressivity might hinder the evolution of structure. We introduced a pressure for expressivity into the experimental model via a highly restrictive filter against ambiguity, which prevented languages to become degenerate (i.e., it guaranteed that each sentence corresponded to a single scene): every time a participant used the same description for more than one meaning (thus introducing homonymous descriptions), they were warned and asked to provide an alternative description. This and similar artificial filters against ambiguity (Carr et al. 2016; Kirby et al. 2008; Silvey et al. 2015; Verhoef 2012) have been previously used as an analogue of a pressure to be expressive which comes from the need to communicate accurately in natural language use. Nonetheless, with a complex meaning space where the discriminating features of meanings might not be clear to the participant and in the absence of a goal to communicate meanings successfully, participants who do not discriminate all meaning features systematically do not have any natural reason to do so in production. The artificial pressure might then force participants to add or delete elements in linguistic expressions which do not necessarily map to any semantics, thereby injecting problematic unconditioned variation into the input for learners at the next generation.

In Experiment 2, we explore whether replacing the artificial pressure against ambiguity with

a more naturalistic mechanism — communicative interaction — facilitates the evolution of complex compositional structure.

4 Experiment 2: transmission and communication

In Experiment 2 we utilise the methodology used in Kirby et al. (2015) and Winters et al. (2015) and introduce a more naturalistic pressure for expressivity through the implementation of communicative interaction at each generation in the transmission chains. We run four transmission chains of eight generations each. Each generation of a chain consists of a pair of participants who are trained on an artificial language based on some linguistic data, and then use that language to communicate with each other, producing a set of linguistic data which provides the input for the next pair of participants in the chain.

4.1 Method

4.1.1 Participants. Sixty-four participants were recruited as per Experiment 1. All participants were native speakers of English (mean age 22 years, age range 18–35); each received a payment of £9. The experiment was conducted in accordance with the ethics procedures of Linguistics and English Language, The University of Edinburgh.

4.1.2 Procedure.

Training and communication regime. In Experiment 2, pairs of participants were asked to individually learn an artificial language which later they would use to communicate with each other. We used the same stimuli as in Experiment 1 and initial languages were generated as per Experiment 1[13].

Experiment 2 was divided into two phases: a training phase and a communication phase. During training, the participants were trained in parallel but separately on a set of 40 out of the 80 scene-description pairings contained in the full language. The two participants in each dyad were trained on the same set of 40 pairings, balanced to contain at least one instance of all meaning features and feature values. They saw each item in the training set three times (order randomised for each participant), giving a total of 120 training trials. After each training trial, participants underwent the same type of recall tests described for Experiment 1: participants were either asked to type in the descriptions they were just presented with, or to select the scene they just saw in the trial.

During communication, the pairs of participants were asked to communicate with each other using the language they had just learned. Pairs communicated the whole set of 80 stimuli during the testing phase, each participant communicating a subset of 40 (again balanced to contain instances of all the different possible values of each feature).There were two roles participants played in this stage, director and matcher. Participants swapped roles at every trial. The director was presented with a scene for 5s (without a description) and then was asked to type in a description for that scene. The description was then sent to their partner, the matcher. The matcher had to identify the scene the director described by selecting a scene out of an array of four displayed in a two by two grid (the target scene and three randomly chosen foils from the remaining 79 scenes in the meaning

[13] We added a few restrictions to the initial languages generated in Experiment 2 that were not present in Experiment 1: we excluded the character $< s >$ in order to avoid its use as a plural marker (as seen in language A3 in Experiment 1) and $< k, q, w, x, y, z >$ were further excluded in order to avoid that participants noticed the restriction. Characters absent in the initial languages were blocked in the keyboard and participants could not enter them in their responses.

space). Full feedback was provided after each trial: participants saw a screen with a red or green background—depending on the communicative success (green for success and red for failure)—which displayed the description the director typed in alongside the meaning the director was trying to convey and the meaning the matcher selected.

Transmission. Pairs of participants were organised into independent transmission chains and the transmission procedure was implemented as per Experiment 1. At each generation, we randomly selected one participant's set of productions out of the pair (composed of 40 scene-description pairings) and used it as the training language for the next generation. Note that the composition of the language transmitted from generation to generation, and therefore the composition of the training set for each participant, differs slightly from Experiment 1: rather than having both participants produce descriptions for all 80 scenes (which would double the duration of the experiment to around 80 2 hours) each participant produces for half of the scenes; we avoid mixing data from multiple participants during transmission since other studies (e.g. Atkinson, Smith, & Kirby in press; Smith et al. 2017) suggest that mixing data in this way at least slows the emergence of regularity.

4.2 Analyses and results

4.2.1 Qualitative analysis. As for languages in Experiment 1, we extracted a matrix of associations between lexical items and their referents in the scenes as explained in section 2.2.2. Figure 9 shows the different lexical items that form the lexicon of the example language C2 (where the C stands for *Communication*, contrasting with the Artificial pressure used in Experiment 1, and 2 indicates the chain number). Examples showing the arrangement of morphs within descriptions are provided in (3).

(3) a. roji ref tube evoto ref
square group slide circle.group group
'A group of squares slid towards a group of circles'

b. evoto ref tube tube roji
circle.group group slide.ongoing square
'A group of circles slide towards a square back and forth'

c. roji babatube babatube evo
square bounce.ongoing circle
'A square bounces towards a circle back and forth'

In Tables 4 and 5 we show the nominal and verbal morphology from all final languages in Experiment 2. As in Experiment 1, Shape and Number are encoded within a complex nominal (see Table 4). Plurality is expressed via full reduplication (C1), free morphs (C2), and suffixation (C2, C3 and C4). All of the languages except for C4 mark *ongoing* Aspect of an event via full reduplication. In C1 and C3 it is the marker for a *terminated* event that is reduplicated, whereas in C2 we observe the full reduplication of the forms encoding Motion. Word order is fixed across languages: focal arguments precede anchor arguments consistently (see section 4.2.4).

Unexpectedly, we observe that half of the languages are underspecified: languages C1 and C4 are underspecified for Motion —i.e., they do not distinguish between *bounce* and *slide*. Although underspecified, the languages are highly systematic and participants' communicative accuracy scores are high: for the languages shown in Tables 4 and 5, dyads communicate successfully

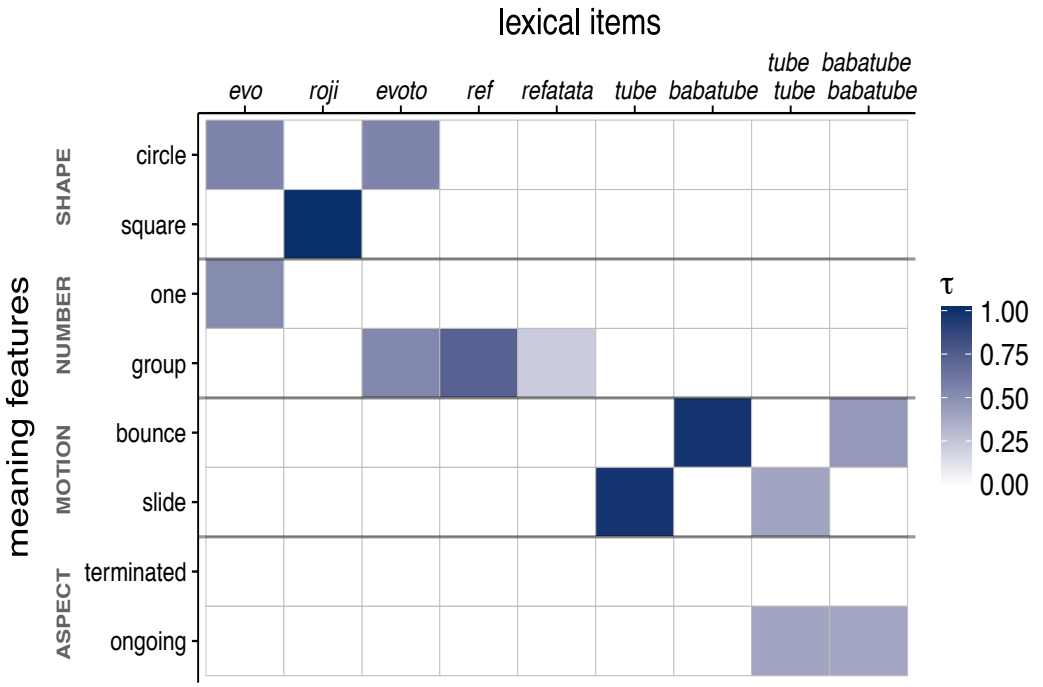


Figure 9. Heatmap showing lexical items in language C2. From left to right, we find that there are two nominal lexical items associated with Shape: *roji* ('square') and *evo* ('circle'). Plurality is generally marked by the morph *refatata* or its clipped equivalent *ref* after nominals to form a complex nominal (e.g., *roji ref* or *roji refatata*, 'a group of squares'). Nevertheless, *evo* takes also a bound morph *-to* as well as the marker *ref* to form the plural 'a group of circles' (i.e., *evoto ref*). We also observe two verbal elements associated with Motion: *babatube* ('bounce') and *tube* ('slide'). Their default Aspect is *terminated* if they appear on their own, and ongoing aspect is marked by full reduplication of the verbal elements: *babatube babatube* ('ongoing bouncing') and *tube tube* ('ongoing sliding').

		C1	C2	C3	C4
circle	one	po	evo	to/ce-	domo
	group	popo	evoto ref	cecede/ceci-	domoge
square	one	vahu	roji	me/me-	pira
	group	vahuvahu	roji ref	mecede/meci-	pirage

Table 4

Nominals within final languages in Experiment 2. Nominals in Language C3 can be expressed through free morphs as well as bound roots, separated by a slash in this table, which take on suffixes marking Aspect. Only free morphs can appear as anchor arguments, whereas both free morphs and bound roots can appear in focal arguments.

		C1	C2	C3	C4
slide	ongo.	jiji-F A	F tube A	F-jijiju mu A	F refugo A
	term.	ji-F A	F tube tube A	F-jiju mu A	F lefugo A
bounce	ongo.	jiji-F A	F babatube A	F-jijiju ju A	F refugo A
	term.	ji-F A	F babatube babatube A	F-jiju ju A	F lefugo A

Table 5
Verbal elements of the final languages in Experiment 2. Focal (F) and Anchor (A) indicate the most frequent position of the nominal arguments in a description.

$\geq$ 95% of the time (see supplementary materials B). Participants thus communicate successfully without encoding all meaning features linguistically; since the foils in discrimination arrays were selected randomly in all matching trials, in most cases, only specifying the shape and number of focal and anchor objects would have been sufficient to disambiguate. Our natural pressure for expressivity is therefore in practice less strict than the artificial pressure used in Experiment 1.

4.2.2 Compositional structure. Figure 10a shows the structure scores obtained in Experiment 2. We performed a segmented linear mixed-effects model with a breakpoint at generation 4 (obtained as per Experiment 1) to explore the effect of generation on linguistic structure, but this time across Experiments 1 and 2. We will call this Model 4. As fixed effects we entered Experiment (Experiment 1 and Experiment 2), Generation, the interaction between Generation and Indicator, and the interaction between Generation, Indicator and Experiment. For all models reported hereafter, we use simple contrast coding for the fixed effect Experiment be-

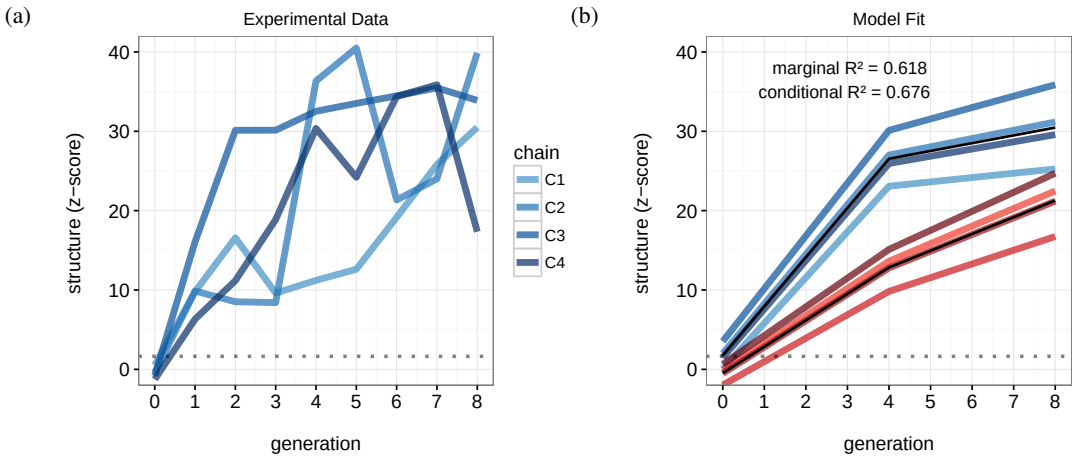


Figure 10. (a) Linguistic structure over generations for each of the four transmission chains in Experiment 2. Linguistic structure increases as languages are transmitted through generations of learners. (b) Fitted values from the mixed-effects regression Model 4 for the four transmission chains in Experiment 1 (red) and the four transmission chains in Experiment 2 (blue). Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas the black lines represent the fixed effects estimates for each experiment. In both plots, the dotted horizontal line represents the lower bound on chance level (z-score 95%CI = $\pm$1.645, one-tailed).

	Generation							
	1	2	3	4	5	6	7	8
chain C1	1	1	1	1	1	1	1	1
chain C2	1	1	0.99	1	1	1	0.96	0.99
chain C3	1	1	1	1	1	1	1	1
chain C4	1	1	1	1	1	1	1	1

Table 6
Proportion of adjacent Shape and Number morphology for a specific object, either focal or anchor. Shape and Number are always encoded adjacent to each other in all languages from the first generation onwards.

cause we are interested in testing main effects (i.e., consider all levels of our categorical predictor Experiment in testing Generation) as well as in comparing levels directly to each other. The intercept in Model 4 is then the grand mean of Experiment 1 and Experiment 2, and we compare Experiment 2 to Experiment 1. As random effects, we introduced intercepts for Chain as well as by-Chain slopes for the effect of Generation. Figure 10b shows the fixed and random estimates obtained in Model 4. The model intercept indicates that languages were highly structured by generation 4 ($\beta = 19.689, SE = 1.931, p < 0.001$). A significant effect of Experiment ($\beta = 6.850, SE = 1.931, p = 0.014$) suggests that languages in Experiment 2 were significantly more structured at generation 4 than languages in Experiment 1. We found a significant effect of Generation ($\beta = 4.763, SE = 0.727, p < 0.001$) and a marginally significant effect of the interaction between Generation and Experiment ($\beta = 1.431, SE = 0.727, p = 0.053$), suggesting that although structure increased in the first four generations across experiments, the increase over generations was marginally greater in Experiment 2. We also found a significant interaction between Generation and Indicator ($\beta = -3.216, SE = 1.275, p = 0.014$) and no effect of the interaction between Generation, Indicator and Experiment ($\beta = -1.994, SE = 1.275, p = 0.123$), suggesting that structure increased more slowly in the second half of transmission chains (i.e. generations 4–8) in both experiments[14]. This suggests a scenario where languages become more stable as a result of the cumulative increase in structure, which facilitates language learning and slows the further development of structure, which is confirmed by an analysis of learning error (see supplementary material A). Altogether, these results suggest that structure increases by generation across experiments, and that it increases more in the first half of the transmission chains. They also suggest that languages in Experiment 2 become more structured faster.

4.2.3 Hierarchical constituent structure: the emergence of complex nominal constituents. Table 6 shows the relative frequency with which morphs encoding Shape and Number appear next to each other in a language. In all four chains, we consistently find complex nominal constituents already by the first generations. As described in section 4.2.1, either Number was marked via reduplication, or morphs encoding Number followed morphs encoding Shape. As in Experiment 1, we tested whether these complex nominals in fact constitute a syntactic category. Figure 11a shows the distributional distance of nominal constituents. All final languages in Experiment 2 obtain z-scores below chance level (-1.645) and thus we conclude that nominal syntactic

[14]Note that although Model 1 did not suggest a steeper slope for structure in the first half of chains in Experiment 1, Model 4 suggests that, looking across both experiments, the increase in structure is more rapid in the first half of the chains.

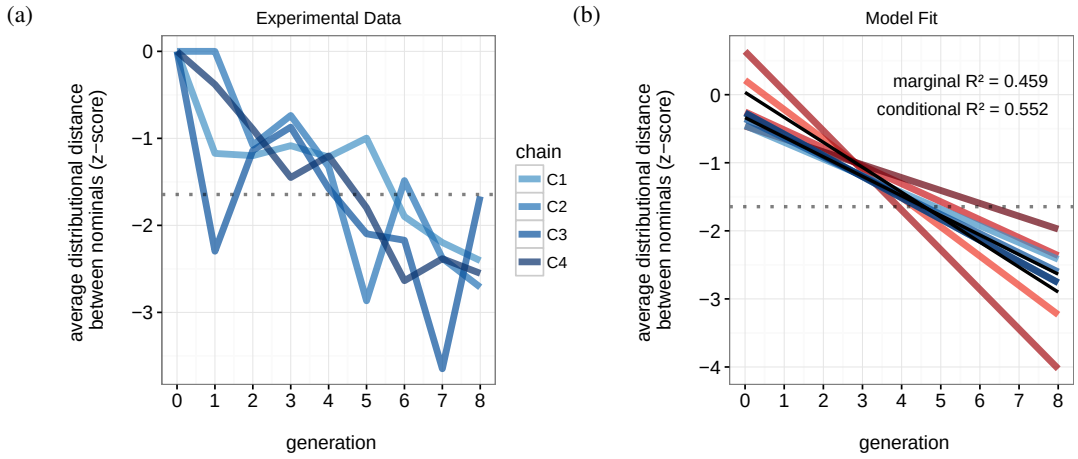


Figure 11. (a) Distributional distance between a language's nominals through generations for each of the four chains in Experiment 2. The dotted line represents the chance level (z-score 95%CI $= \pm 1.645$, one-tailed); z-scores below it indicate that the distributional similarity between nominals is unlikely to arise by chance. Distributional distance between nominals decreases with generation as languages become more structured, suggesting the emergence of a nominal syntactic category. We observe that all the average distance between nominals are below chance level in the last two generations. (b) Fitted values from Model 5 for the four transmission chains in Experiment 1 (red) and the four transmission chains in Experiment 2 (blue). Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas the black lines represent the fixed effects estimates for each experiment.

categories evolve via cultural transmission. We also performed a linear mixed-effects model, which we will call Model 5, to test the effect of generation on the distributional distance across Experiments 1 and 2. As fixed effects, we entered Generation (centred) and Experiment (Experiment 2 vs. Experiment 1) as well as their interaction. As random effects, we introduced an intercept for Chain and a by-Chain random slope for the effect of Generation. Figure 11b shows the fitted values of Model 5 for fixed and random effects. Results showed a significant effect of Generation ($\beta = -0.327, SE = 0.061, p < 0.001$) and no significant interaction between Generation and Experiment ($\beta = 0.040, SE = 0.061, p = 0.5371$) , suggesting that the distributions in which nominals appear became more similar by generation to a comparable degree across experiments. There was no effect of Experiment ($\beta = -0.027, SE = 0.115, p = 0.819$), suggesting that both experiments obtained similar estimates at the intercept[15].There is a clear difference between experiments in the amount of languages in which we observe the evolution of a nominal syntactic categories: only in two in Experiment 1 and in all four in Experiment 2. Nevertheless, we observe that distance between nominals diminishes by generation across all languages in both experiments.

[15]It is worth noting that larger z-scores can be obtained from languages with larger lexicons because the probability of selecting morphs that are nominals during repeated random sampling is lower (see section 2.2.3). Because languages in Experiment 1 have larger lexicons (primarily because they are not as structured and systematic as languages in Experiment 2), z-scores obtained in the two languages that evolved nominal syntactic categories are lower (see section 2.3.3), leading to an average z-score similar to that in Experiment 2.

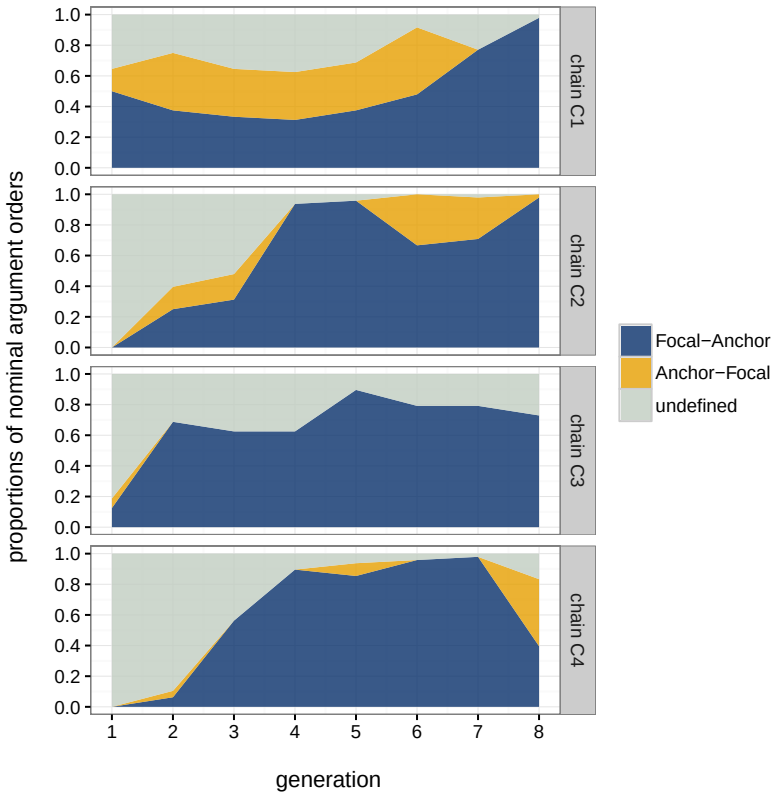


Figure 12. Proportion of word order types for nominal arguments by chain and generation. A fixed focal-anchor word order evolves across chains.

4.2.4 Word order rules for nominal arguments. The stacked area graphs in Figure 12 show the proportions of focal-anchor, anchor-focal and undefined orders at each generation for each of the four transmission chains in Experiment 2. We observe that the proportion of undefined order decreases as languages are transmitted through generations, and the proportion of focal-anchor orders increases rapidly. Moreover, we observe that the order of nominal arguments is mostly fixed in the last few generations. Word order regularity shows yet another aspect in which the languages that evolved in Experiment 2 are more systematic than those in Experiment 1.

We ran a linear mixed-effects model, which we will call Model 6, to test the effect of generation on the variability of nominal argument orders in languages in Experiments 1 and 2 —calculated by the entropy of the system of orders as described in section 2.2.4. We entered Generation (centred), Experiment (Experiment 2 vs. Experiment 1) and their interaction as fixed effects; as random effects we entered intercepts for Chain as well as by-Chain slopes for the effect of Generation. Figure 13 shows the nominal order variability scores of the experimental data (Figure 13a) as well as the fitted values of Model 6 for fixed and random effects (Figure 13b). Results show a significant effect of Generation ($\beta = -0.073, SE = 0.014, p < 0.001$) and no significant interaction between Generation and Experiment (β=$-0.005, SE = 0.014, p = 0.744$), suggesting that entropy decreases by generation to a similar degree across experiments, and therefore that the order of nominal arguments becomes more consistent as languages are transmitted through generations of learners. We

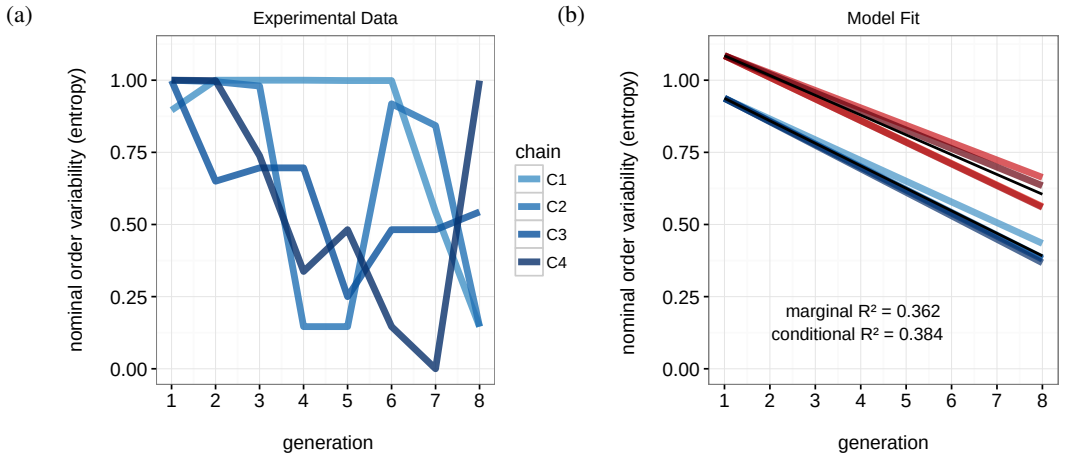


Figure 13. (a) Variability of nominal argument orders by generation and chain in Experiment 2. (b) Fitted values from the mixed-effects regression Model 6 for the four transmission chains in Experiment 1 (red) and the four transmission chains in Experiment 2 (blue). Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas the black lines represent the fixed effects estimates for each experiment.

also found a significant effect of Experiment ($\beta = -0.090, SE = 0.034, p = 0.022$) indicating that the order of nominals is less consistent in the languages in Experiment 1 at the intercept.

5 Discussion

5.1 The evolution of complex compositional structure

In the present study we explored whether and how complex compositional structure evolves through cultural transmission. We ran two experiments in which we manipulated the nature of a pressure for expressivity (artificial vs. communication) whilst keeping constant a learning bottleneck in transmission, which promotes the need for generalisation and thus for the learnability of the language. We showed that compositional structure evolved from the trade-off between these two pressures across experiments thus replicating the results found in previous IALL studies (Beckner et al. 2017; Kirby et al. 2008, 2015) but with a more complex meaning space. Initial unstructured languages, where each scene in the meaning space was mapped to a randomly generated string of characters, became structured as they were transmitted down generations of participants: languages developed one-to-one form-meaning mappings as well as an isomorphic relation between syntactic and semantic structure. The effect of generation on linguistic structure was not equally pronounced across generations: structure developed more rapidly initially and more slowly later in the chains, as languages became more stable as a result of the cumulative increase in structure, which facilitates language learning. These non-linear evolutionary trajectories are reminiscent of the (more or less pronounced) logarithmic curves shown in many models of cultural evolution (Boyd & Richerson 1988; Claidière & Sperber 2007; Griffiths, Kalish, & Lewandowsky 2008; Henrich & Boyd 2002; Mesoudi 2011).

The main contribution of the present study is the evidence for the cultural evolution of the complexity in linguistic structure: the emergent morphosyntactic properties across these languages

mimic the processes and properties of natural languages more closely than previously attested in IALL studies. This study has demonstrated that, in combination with a sufficiently complex meaning space, compositional hierarchical constituent structure (i.e., where complex meaningful units composed of meaningful units which are complex themselves) can evolve from the trade-off between the (sometimes) competing pressures at play in cultural transmission, learnability and expressivity. Moreover, the experimental results suggest that the same trade-off can lead to the emergence of "positional" compositionality (Galantucci & Garrod 2011). We showed the emergence of word order rules by which the position in which nominal constituents appear determines the semantic role they perform in the motion event. In comparison to other strategies such case-marking, word order rules also help language users to minimise the number of morphs that need to be memorized to unambiguously convey a given meaning. The emergence of compositional hierarchical structure and word order rules relaxes the constraints on productivity of basic compositionality (e.g. Kirby et al. 2008, 2015) mirroring the type of structure found in natural languages more closely, and providing an experimental deomstration of the second part of the definition of compositionality (i.e,"the meaning of the whole is determined by the meaning of its parts and *the way they are combined*").

The languages that evolved structure comply with the characteristics of configurational languages where word order is fairly fixed and sentences are mainly composed of morphosyntactically continuous expressions (i.e., continuous constituents, without long-distance dependencies). Moreover, more frequent and salient meaning features such as Shape (Gentner 1982; Landau, Smith, & Jones 1988) and its associated feature Number were always encoded, unlike Motion and Aspect which were not always both encoded. Although the strategy for marking number varied across the evolved languages in this study (e.g., simulfixation, suffixation, reduplication, suppletion or plural word), morphs encoding Shape and Number always appeared adjacently, forming continuous nominal constituents. This is consistent with the universal tendency (seen in English) to mark nominal number in the noun or in its immediate periphery (Dryer 2013a). A systematic nominal category also emerged relatively early on in the chains. Although this might be due to type frequency, it conforms to a noun bias parallel to that suggested in language acquisition in many languages (Dhillon 2010). Moreover, it is also worth noting that where fixed word order evolved, regardless of the position of verbal elements within the system (i.e., whether it appeared sentence medial or final), focal objects always precede anchor objects, consistent with the universal Agent-first tendency in natural languages (including English) (Dryer 2013b; Greenberg 1966), and a universal processing bias (Gibson 2000; Hawkins 2004; Marantz 2005).

5.2 The effect of communicative interaction in the evolution of complex compositional structure

Although similar linguistic structure evolved across experiments, it did so to varying degrees determined by the nature of the pressure for expressivity, that is, either an artificial pressure against ambiguity in production or communicative interaction. Results show that the substitution of an artificial pressure for communicative interaction eases the evolution of linguistic structure (for a similar conclusion, see Carr et al. 2016). With the inclusion of communicative interaction, structure emerged more rapidly, and languages become significantly structured by the first generation. Moreover, all languages in Experiment 2 evolve to be significantly structured and systematic: unlike in Experiment 1, descriptions only contain morphology which has a semantic mapping to the constituent parts of the meaning they describe and word order is fixed across *all* four languages. The only aspect in which languages that evolved in Experiment 1 are more systematic than those in

Experiment 2 is in the sublexical structure within morphology encoding Shape: languages in Experiment 2 did not evolve nominal category markers and thus, string-similarity between nominals is lower within languages in Experiment 1 (e.g., between *ron* and *mon*).

These differences observed between conditions suggest that the expressivity that communication promotes is not analogous to an artificial pressure against ambiguity—at least given the presence of a complex meaning space and the highly restrictive artificial pressure we implemented. Half of the languages that evolved in Experiment 2 were underspecified for one meaning feature (either Motion or Aspect, but never Shape or Number). This suggests that provided that not all meaning features are required to be discriminated at every communicative event, communicative interaction does not impose as strong a pressure for expressivity as assumed in Experiment 1. Contrary to what we assumed in constructing our artificial pressure against ambiguity, participants were often not required to discriminate all features of the meanings for communication to be successful. It is not necessary (or logically possible) to specify all aspects of a meaning in a concrete communicative event—be it because they are provided by the context or they are simply not required; therefore, it is more economical or at least sufficient to encode the minimum meaning features, minimising the effort of unambiguously conveying a message (Brochhagen, Franke, & van Rooij 2016; Winters et al. 2015; Winters, Kirby, & Smith 2018). The differences between the underspecification found in Experiment 2 compared to the full expressivity found with a similar design in Kirby et al. (2015) is most probably due to the differing complexity of the meaning space and the size of the context array the matcher has to select meanings from during communication. Whilst in Kirby et al. (2015) participants are asked to select an object out of a context array of 6 with a much simpler meaning space (i.e., 12 objects in total, only differing in shape and fill-pattern), participants in Experiment 2 are only asked to discriminate the scene conveyed by the partner out of an array of four (randomly selected) at each communication trial and with a substantially more complex meaning space (i.e., 80 meanings, with 5–7 features each). The probability of having to discriminate every single meaning feature value of a scene is lower in our design than it is in Kirby et al. (2015).

Altogether, these results suggest that a coordination pressure in communicative interaction contributes significantly to the emergence of linguistic structure. The possibility of coordination between participants leads to a very early emergence of structure in Experiment 2 (see also Raviv, Meyer, & Lev-Ari 2018; Winters et al. 2018). With a shared goal to communicate accurately, participants might prioritise the establishment of conventions with partners to bootstrap communication—even at the expense of faithful reproduction of the learned language. It is probable that the inclusion of communication and thus of the explicit goal of arriving at a shared system for communication results in conscious design by language users more than in Experiment 1. Nevertheless, the degree of structure increases as languages are transmitted through a learning bottleneck and thus, similarly to Theisen, Oberlander, and Kirby (2010) and Theisen-White et al. (2011) in the graphic modality, results show that a certain degree of structure can emerge during communicative interaction but it is through iterated learning that it accumulates (see also Raviv et al. 2018 for a similar cumulative effect with an expanding lexicon and a turnover of communicative partners) .

In sum, the addition of communicative interaction to transmission facilitates the evolution of linguistic structure. The effect of communication in this study is thus not reducible to a pressure against ambiguity in production. The coordination pressure at play during communication facilitates the conventionalisation of lexical items and grammatical rules (Garrod & Anderson 1987; Lewis 1968). Moreover, communicative interaction (i.e., without a requirement of full discrimination at each communicative event) does not impose such a hard constraint on expressivity as

assumed in our artificial pressure against ambiguity: most languages are underspecified for one meaning feature. This can be explained in terms of effort minimisation in communication given that underspecification minimises effort in production, and communicative effectiveness was not compromised under the communicative context provided.

6 Conclusion

Human communication systems across the world posses a remarkably productive linguistic structure. This productivity is facilitated by (at least) two properties: hierarchical constituency and compositionality. One explanation for the existence of these properties appeals to the idea that linguistic structure evolved to adapt to the selective pressures present in cultural transmission: learnability and expressivity. Our study has explored the hypothesis that, with a complex enough world to communicate about, the behavioural product of the mechanisms operating in individuals during language learning and production can lead to the evolution of hierarchical compositional structure over cultural time, as suggested in simulations by Kirby (2002) and Batali (2002). Our findings reveal that when pressures to generalise and to be expressive are both present, languages become hierarchically compositional, which maximises the language's learnability without jeopardising its expressivity. Moreover, our study suggests that the trade-off between lernability and expressivity pressures is not only present intergenerationally in cultural transmission but also within a single generation during communicative interaction (Raviv et al. 2018; Winters et al. 2018). In comparison to an artificial pressure against ambiguity during individual production, the inclusion of communicative interaction facilitates the evolution of linguistic structure. More generally, this study provides support for the claim that cultural transmission is a linking mechanism by which the advantages provided by compositional hierarchical structure in solving the immediate problems confronting individual learner/user lead to such structures permeating language.

References

Atkinson, M., Smith, K., & Kirby, S. (in press). Adult learning and language simplification. *Cognitive Science*.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge: Cambridge University Press.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 111–172). Cambridge, UK: Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, lzx001. doi: 10.1093/jole/lzx001

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*(3), 177–226. doi: 10.1016/j.plrev.2005.06.001

Brochhagen, T., Franke, M., & van Rooij, R. (2016). Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (bayesian) learning and functional selection. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2081–2086). Austin, TX: Cognitive Science Society.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2016). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive science*, *41*(4), 892–923. doi: 10.1111/cogs.12371

Chomsky, N. (1957). *Syntactic structures*. Berlin: De Gruyter.

Chomsky, N. (1965). *Aspects of theory of syntax*. Cambridge, MA: The MIT Press.

Chomsky, N. (1980). Rules and representations. *Behavioral and brain sciences*, *3*(01), 1–15. doi: 10.1017/S0140525X00001515Published

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and brain sciences*, *31*(5), 489–509. doi: 10.1017/S0140525X08004998

Claidière, N., & Sperber, D. (2007). The role of attraction in cultural evolution. *Journal of Cognition and Culture*, *7*(1), 89–111. doi: 10.1163/156853707X171829

Dhillon, R. (2010). Examining the 'noun bias': A structural approach. *University of Pennsylvania Working Papers in Linguistics*, *16*(1), 7.

Dryer, M. S. (2013a). Coding of nominal plurality. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/chapter/33`

Dryer, M. S. (2013b). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/chapter/81`

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5*, 11. doi: 10.3389/fnhum.2011.00011

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. doi: 10.1016/0010-0277(87)90018-7

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. technical report no. 257.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 95–126. Retrieved from `https://tedlab.mit.edu/tedlab_website/researchpapers/Gibson_2000_DLT.pdf`

Greenberg, J. H. (1966). *Universals of language* . MIT press.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3503–3514. doi: 10.1098/rstb.2008.0146

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, *29*(2), 147–160. doi: 10.1002/j.1538-7305.1950.tb00463.x

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Henrich, J., & Boyd, R. (2002). On modeling cognition and culture: Why cultural evolution does not require replication of representations. *Journal of cognition and culture*, *2*(2), 87–112. doi: 10.1163/156853702320281836

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–111. doi: 10.1038/scientificamerican0960-88

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. doi: 10.2307/2332226

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, *33*(3), 239–251. doi: 10.1093/biomet/33.3.239

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford: OUP Oxford.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110. doi: 10.1109/4235.918430

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–204). Cambridge, UK: Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. doi: 10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer. doi: 10.1007/978-1-4471-0663-0_6

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. doi: 10.1016/j.cognition .2015.03.016

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). *Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). r package version 2.0-3.* Retrieved from `https://cran.r-project.org/web/packages/lmerTest/index.html`

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321. doi: 10.1016/0885-2014(88)90014-7

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710). Retrieved from `http://adsabs.harvard.edu/abs/1966SPhD...10..707L` (Provided by the SAO/NASA Astrophysics Data System)

Lewis, D. (1968). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*(2-4), 429–445. doi: 10.1515/tlir.2005.22.2-4.429

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The cappuccino model. In *Proceedings of the cognitive science society* (Vol. 33). Austin, TX. Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.208.1404&rep=rep1&type=pdf`

Mesoudi, A. (2011). Variable cultural acquisition costs constrain cumulative cultural evolution. *PloS one*, *6*(3), e18239. doi: 10.1371/journal.pone.0018239

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3), 373–398. doi: 10.1111/j.1755-2567 .1970.tb00434.x

Nowak, I., & Baggio, G. (2016). The emergence of word order and morphology in compositional languages via multigenerational signaling games. *Journal of Language Evolution*, 137–150.

doi: 10.1093/jole/lzw007

Pagin, P. (2012). Communication and the complexity of semantics. In *The oxford handbook of compositionality.* Oxford University Press.

Pagin, P. (2013). Compositionality, complexity, and evolution. In *Proceedings: Symposium on language acquisition and language evolution* (pp. 51–62). Department of Linguistics, Stockholm University. Retrieved from http://www.ling.su.se/polopoly_fs/1.142249.1376032271!/menu/standard/file/Pagin_KVA_Symposium_20130808.pdf

Pagin, P., & Westerståhl, D. (2010). Compositionality i: Definitions and variants. *Philosophy Compass*, *5*(3), 250–264. doi: 10.1111/j.1747-9991.2009.00228.x

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775–793. doi: 10.1111/cogs.12102

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305. doi: 10.3758/MC.36.7.1299

Pullum, G. K., & Scholz, B. C. (2007). Systematicity and natural language syntax. *Croatian Journal of Philosophy*, *7*(21), 375–402. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.7486

R Core Team. (2000). *R language definition.* Available from CRAN sites. Retrieved from cran.r-project.org

Raviv, L., Meyer, A., & Lev-Ari, S. (2018). Compositional structure can emerge without generational transmission. *Cognition*. doi: 10.1016/j.cognition.2018.09.010

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, *87*, 237. doi: 10.1111/b.9781118301753.2015.00013.x

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, *14*(9), 411–417. doi: 10.1016/j.tics.2010.06.006

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive science*, *39*(1), 212–226. doi: 10.1111/cogs.12150

Smith, K., & Kirby, S. (2012). Compositionality and linguistic evolution. In *The oxford handbook of compositionality.* Oxford University Press.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Phil. Trans. R. Soc. B*, *372*(1711), 20160051. doi: 10.1098/rstb.2016.0051

Szabó, Z. G. (2012). The case for compositionality. In *The oxford handbook of compositionality.* Oxford University Press.

Theisen, C. A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11*(1), 14–32. doi: 10.1075/is.11.1.08the

Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 956–961). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.208.1489

Vanhove, J. (2014, 20th August). *Calibrating p-values in 'flexible' piecewise regression models.* Retrieved from https://janhove.github.io/analysis/2014/08/20/adjusted-pvalues-breakpoint-regression.

van Trijp, R. (2012). The evolution of case systems for marking event structure. In L. Steels (Ed.), *Experiments in cultural language evolution* (Vol. 3, pp. 169–205). Amsterdam: John

Benjamins.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, *4*(4), 357–380. doi: 10.1515/langcog-2012-0019

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(03), 415–449. doi: 10.1017/langcog.2014.35

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: The MIT Press.

Zuidema, W. H. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in neural information processing systems* (pp. 51–58). Retrieved from `http://machinelearning.wustl.edu/mlpapers/paper_files/CS07.pdf`

Supplementary material A
Learning error

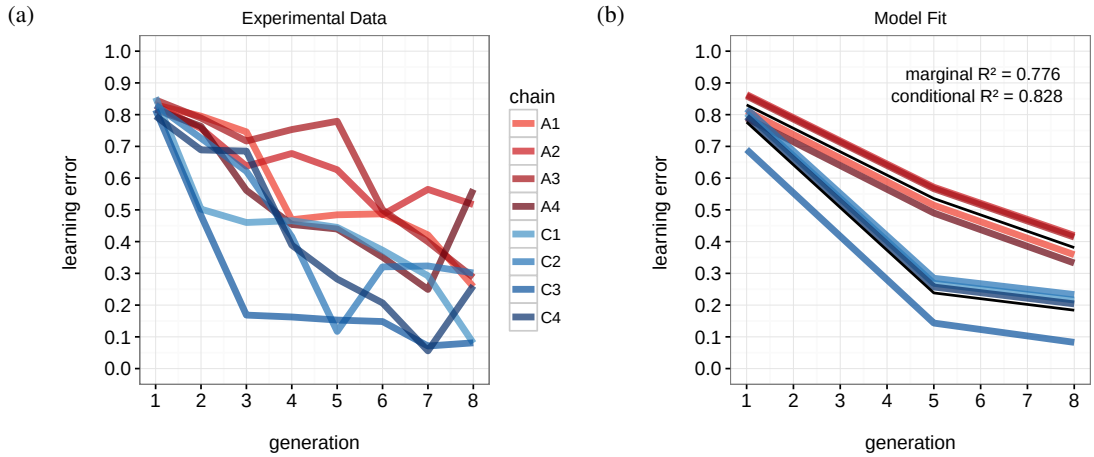


Figure A1. (a) Learning error over generations across experiments for each of the transmission chains. Whilst in Experiment 1 (Artificial) and Experiment 2 (Communication) learning error decreases consistently as languages are transmitted through generations of learners. (b) Fitted values from the mixed-effects for Experiments 1 and 2. Coloured lines represent the random slopes estimates (for generation) depending on random intercepts (individual chains), whereas the black lines represent the fixed effects estimates for each experiment.

In order to show that the increase in the compositional structure shown in the experiments above is shaped by the need for language to be learnable and thus favours the learnability of the languages, in this section we show that languages' learnability does indeed increase by generation.

Following the measures used in Kirby et al. (2008, 2015) to evaluate learnability, we define and increase in learnability by the decrease of learning error from the learned system to the produced system. In order to quantify the learning error at each generation, we computed the average normalised Levenshtein edit-distance (LD) (Levenshtein 1966) between the descriptions produced at generation *g* and those produced at generation *g-1* to refer to the same scenes; we normalised the distances such that the maximum error is 1. Figure A1a shows the learning error in Experiments 1 and 2. We observe a decrease in learning error across generations; however, in Experiment 2, this decrease is greater in the first generations and learning error is lower by the final generation.

We performed a segmented linear mixed-effects model with a breakpoint at generation 5 to explore the effect of generation on learning error across experiments. As fixed effects we entered Generation with an Indicator variable nested and Experiment, as well as their interaction. We used simple contrast coding for Experiment so the intercept is the grand mean across experiments, and we compared Experiment 1 to Experiment 2. As random effects we entered intercepts for Chain as well as by-Chain random slopes for the effect of Generation. We found a significant effect of Generation ($\beta = -0.104, SE = 0.010, p < 0.001$) as well as a significant interaction between Generation and Indicator (β=0.069, SE = 0.021, p=0.002), suggesting that learning error decreased significantly by generation but less in the second half of the chain as languages become more learnable and thus stable. We also found a significant effect of Experiment ($\beta = 0.149, SE = 0.030, p < 0.001$), suggesting that learning error at generation 5 was significantly higher in Experiment 1 compared

to Experiment 2. Moreover, we obtained significant interactions between Generation and Experiment ($\beta = 0.030, SE = 0.010, p = 0.004$) and between Generation, Experiment and Indicator ($\beta = -0.047, SE = 0.021, p = 0.033$), suggesting that, compared to Experiment 2, the decrease in learning error in Experiment 1 was lower in the first half of the transmission chains but it did not abate after generation 5, which at the same time indicated that languages in Experiment 1 were not established in the last generations as languages in Experiment 2 were.

Supplementary material B
Underspecification and communicative accuracy in Experiment 2

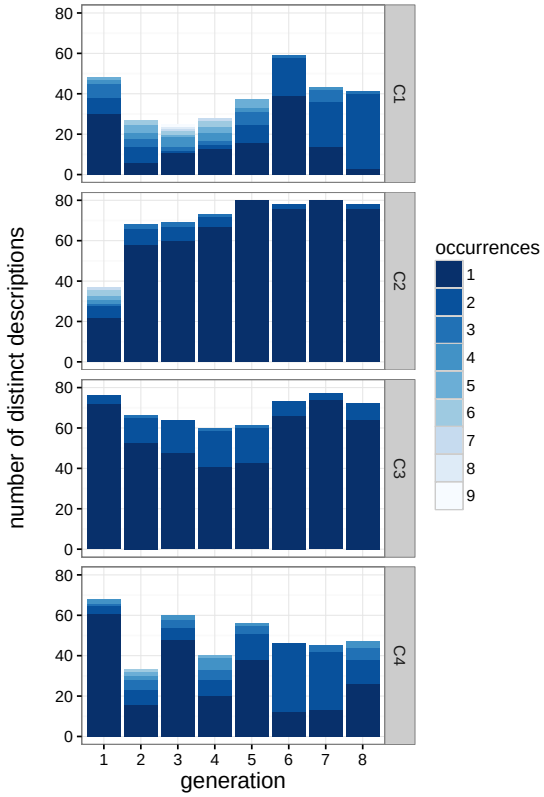


Figure B1. Number of distinct descriptions and their occurrences in a language. A fully expressive language would have 80 distinct descriptions, one per scene. Any description that occurs more than once in a language introduces ambiguity into the system. Languages C2 and C3 at the final generations are fairly expressive, most of their descriptions only occur once and thus are only associated with one scene. By contrast, the final languages C1 and C4 are underspecified and thus less expressive: most of the descriptions are homonyms often corresponding to two different scenes (i.e., corresponding to the observed underspecification of either Motion or Aspect meaning features).

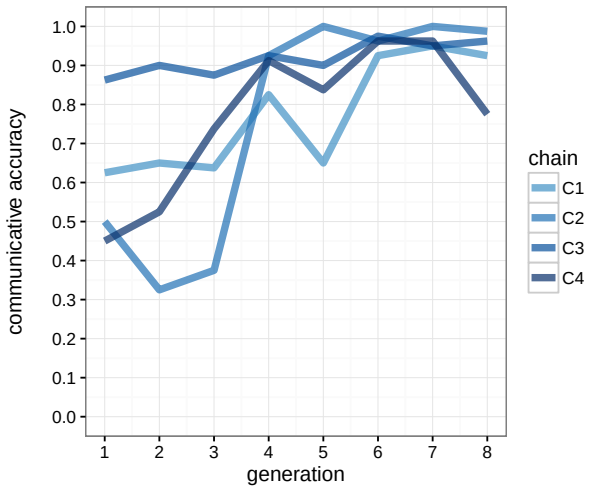


Figure B2. This graph shows the communicative accuracy as a proportion of the successes during communication (80 trials) between pairs of participants in Experiment 2. We observe an increase in communicative accuracy in the first five generations, where it is on average $\hat{p} > 0.8$. It later continues on increasing and it is $\hat{p} > 0.9$ in all languages by the final generations. Exceptionally, communicative accuracy drops to $\hat{p} = 0.775$ in the last generation of chain C4, where we also observed a drop in linguistic structure.